A Bregman Extension of quasi-Newton updates II: Convergence and Robustness Properties

Takafumi Kanamori
Nagoya University
kanamori@is.nagoya-u.ac.jp

Atsumi Ohara
Osaka University
ohara@sys.es.osaka-u.ac.jp

Abstract

We propose an extension of quasi-Newton methods, and investigate the convergence and the robustness properties of the proposed update formulae for the approximate Hessian matrix. Fletcher has studied a variational problem which derives the approximate Hessian update formula of the quasi-Newton methods. We point out that the variational problem is identical to optimization of the Kullback-Leibler divergence, which is a discrepancy measure between two probability distributions. Then, we introduce the Bregman divergence as an extension of the Kullback-Leibler divergence, and derive extended quasi-Newton update formulae based on the variational problem with the Bregman divergence. The proposed update formulae belong to a class of self-scaling quasi-Newton methods. We study the convergence property of the proposed quasi-Newton method, and moreover, we apply the tools in the robust statistics to analyze the robustness property of the Hessian update formulae against the numerical rounding errors included in the line search for the step length. As the result, we found that the influence of the inexact line search is bounded only for the standard BFGS formula for the Hessian approximation. Numerical studies are conducted to verify the usefulness of the tools borrowed from robust statistics.

1 Introduction

We consider quasi-Newton methods for the unconstrained optimization problem

$$\text{minimize } f(x), \quad x \in \mathbb{R}^n, \tag{1}$$

in which the function $f : \mathbb{R}^n \to \mathbb{R}$ is twice continuously differentiable on $\mathbb{R}^n$. The quasi-Newton method is known to be one of the most successful methods

for unconstrained function minimization. Details are shown in [15, 13] and references therein.

The main purpose of this paper is to present an extended framework of quasi-Newton method, and to study the robustness property of quasi-Newton update formulae against numerical errors of line search. There are mainly two standard quasi-Newton method; one is the DFP formula and the other is the BFGS formula. Fletcher [7] has pointed out that the standard formulae, DFP and BFGS, are obtained as the optimal solution of a variational problem over the set of positive definite matrices. Along this line, we extend the quasi-Newton update formula. Then, we study the robustness property of the extended quasi-Newton methods, where we apply some techniques exploited in the field of robust statistics [11].

We briefly introduce quasi-Newton formulae and its variational result. In quasi-Newton method, a sequence $\{x_k\}_{k=0}^{\infty} \subset \mathbb{R}^n$ is successively generated in a manner such that $x_{k+1} = x_k - \alpha_k B_k^{-1} \nabla f(x_k)$. The coefficient $\alpha_k \in \mathbb{R}$ is a step-size computed by a line search, and B_k is a positive definite matrix approximating the Hessian matrix $\nabla^2 f(x_k)$ at the point x_k. Let s_k and y_k be column vectors defined by

$$s_k = x_{k+1} - x_k = -\alpha_k B_k^{-1} \nabla f(x_k), \qquad y_k = \nabla f(x_{k+1}) - \nabla f(x_k).$$

We need a Hessian approximation B_{k+1} for $\nabla^2 f(x_{k+1})$ to keep on the computation. In the DFP method, B_{k+1} is given by

$$B_{k+1} = B^{DFP}[B_k; s_k, y_k] := B_k - \frac{B_k s_k y_k^\top + y_k s_k^\top B_k}{s_k^\top y_k} + s_k^\top B_k s_k \frac{y_k y_k^\top}{(s_k^\top y_k)^2} + \frac{y_k y_k^\top}{s_k^\top y_k}, \tag{2}$$

and the BFGS method provides the different formula such that

$$B_{k+1} = B^{BFGS}[B_k; s_k, y_k] := B_k - \frac{B_k s_k s_k^\top B_k}{s_k^\top B_k s_k} + \frac{y_k y_k^\top}{s_k^\top y_k}, \tag{3}$$

When $B_k \in \mathrm{PD}(n)$ and $s_k^\top y_k > 0$ hold, both $B^{DFP}[B_k; s_k, y_k]$ and $B^{BFGS}[B_k; s_k, y_k]$ are also positive definite matrices. In practice, the Cholesky decomposition of B_k will be successively updated in order to compute the search direction $-B_k^{-1} \nabla f(x_k)$ efficiently. The idea of updating Cholesky factors is pioneered by Gill and Murray [9]. Note that the equality

$$B^{DFP}[B_k; s_k, y_k]^{-1} = B^{BFGS}[B_k^{-1}; y_k, s_k]$$

holds. Hence, the update formula for the inverse $H_{k+1} = B_{k+1}^{-1}$ can be directly derived from $H_k = B_k^{-1}$ without computing inversion of matrix.

We introduce a variational approach in quasi-Newton methods. Let $\mathrm{PD}(n)$ be the set of all n by n symmetric positive definite matrices, and the function $\psi : \mathrm{PD}(n) \to \mathbb{R}$ be a strictly convex function over $\mathrm{PD}(n)$ defined by

$$\psi(A) = \mathrm{tr}(A) - \log\det A.$$

Fletcher [7] has shown that the DFP update formula (2) is obtained as the unique solution of the constraint optimization problem,

$$\min_{B\in\mathrm{PD}(n)} \ \psi(B_k^{1/2}B^{-1}B_k^{1/2}) \quad \text{subject to} \ \ Bs_k = y_k,$$

where $A^{1/2}$ for $A \in \mathrm{PD}(n)$ is the matrix satisfying $A^{1/2} \in \mathrm{PD}(n)$ and $(A^{1/2})^2 = A$. The BFGS formula is also obtained as the optimal solution of

$$\min_{B\in\mathrm{PD}(n)} \ \psi(B_k^{-1/2}BB_k^{-1/2}) \quad \text{subject to} \ \ Bs_k = y_k,$$

in which $B_k^{-1/2}$ denotes $(B_k^{-1})^{1/2}$ or equivalently $(B_k^{1/2})^{-1}$.

It will be worthwhile to point out that the function ψ is identical to Kullback-Leibler(KL) divergence [1, 12] up to an additive constant. Let $N_n(0, P)$ be the n dimensional Gaussian distribution with mean zero and variance-covariance matrix $P \in \mathrm{PD}(n)$, then the KL-divergence between $N_n(0, P)$ and $N_n(0, Q)$ is defined by

$$\mathrm{KL}(P, Q) = \mathrm{tr}(PQ^{-1}) - \log\det(PQ^{-1}) - n$$

which is equal to $\psi(Q^{-1/2}PQ^{-1/2}) - n$. The KL-divergence is regarded as a generalization of squared distance over the space of probability distributions. Using the KL-divergence, we can represent the update formulas as the optimal solution of the following minimization problems,

$$\text{(DFP)} \qquad \min_{B\in\mathrm{PD}(n)} \ \mathrm{KL}(B_k, B) \quad \text{subject to} \ \ Bs_k = y_k, \tag{4}$$

$$\text{(BFGS)} \qquad \min_{B\in\mathrm{PD}(n)} \ \mathrm{KL}(B, B_k) \quad \text{subject to} \ \ Bs_k = y_k. \tag{5}$$

The KL-divergence is asymmetric, that is, $\mathrm{KL}(P, Q) \neq \mathrm{KL}(Q, P)$ in general. Hence the above problems will provide different solutions.

Here is the brief outline of the article. In Section 2 we introduce the so-called Bregman divergence which is an extension of the KL-divergence. In

Section 3, an extended quasi-Newton formula is derived based on the Bregman divergence. In Section 4, the convergence property of the proposed quasi-Newton method is studied, and Section 5 is devoted to discuss the robustness of the Hessian update formula. Numerical simulations are presented in Section 6. We conclude with a discussion and outlook in Section 7. Some proofs of the theorems are postponed to Appendix.

Throughout the paper, we use the following notations: The set of positive real numbers are denoted as $\mathbb{R}_+ \subset \mathbb{R}$. Let $\det A$ be the determinant of square matrix A, and $\mathrm{GL}(n)$ denotes the set of n by n non-degenerate real matrices. The set of all n by n real symmetric matrices is denoted as $\mathrm{Sym}(n)$, and let $\mathrm{PD}(n) \subset \mathrm{GL}(n) \cap \mathrm{Sym}(n)$ be the set of n by n symmetric positive definite matrices. For two square matrices A, B, the inner product $\langle A, B \rangle$ is defined by $\mathrm{tr}(AB^\top)$, and $\|A\|_F$ is the Frobenius norm defined by the square root of $\langle A, A \rangle$. Throughout the paper we only deal with the inner product of symmetric matrices, and the transposition in the trace will be dropped. For a vector x, $\|x\|$ denotes the Euclidean norm. The first and second order derivative of a function $f : \mathbb{R} \to \mathbb{R}$ are denoted as f' and f'', respectively.

2 Bregman Divergence induced from Potential Functions

As introduced in Section 1, the update formulae of the DFP and the BFGS methods are derived from the optimization problem of KL-divergence. In this section we introduce Bregman divergence [3] which is an extension of the KL-divergence. Especially we focus on the Bregman divergence induced from potential function. Then, we present extended quasi-Newton formulae derived from the variational problem for the Bregman divergence.

Let $\varphi : \mathrm{PD}(n) \to \mathbb{R}$ be a differentiable, strictly convex function that maps positive definite matrices to real numbers. We define *Bregman divergence* of the matrix P from the matrix Q as

$$D(P, Q) = \varphi(P) - \varphi(Q) - \langle \nabla\varphi(Q), P - Q \rangle, \tag{6}$$

where $\nabla\varphi(Q)$ is the n by n matrix whose (i, j) element is given as $\frac{\partial \varphi}{\partial Q_{ij}}(Q)$. The strict convexity of φ guarantees that $D(P, Q)$ is non-negative and equals to zero if and only if $P = Q$ holds. Figure 1 illustrates the relation between the function φ and the Bregman divergence. Note that $D(P, Q)$ is convex in P but not necessarily convex in Q. Bregman divergences have been well studied for nearness problems in the fields of statistics and machine learning [2, 6, 14].

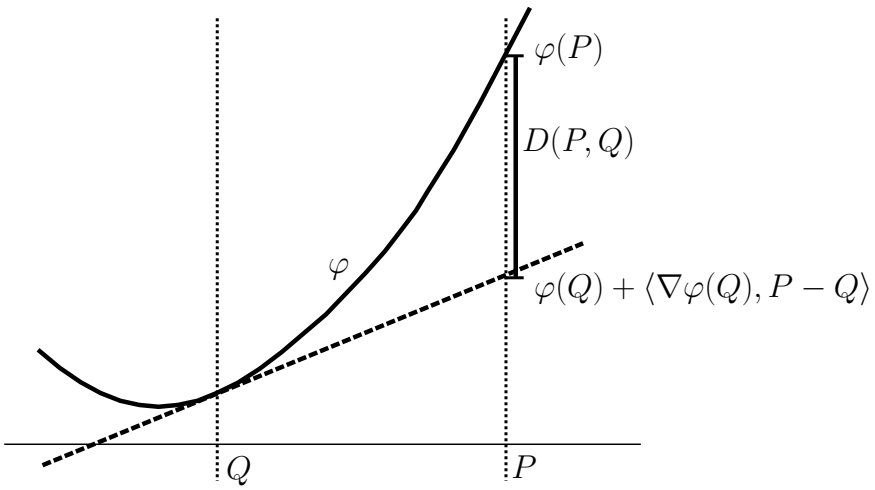

Figure 1: The Bregman divergence defined by the strictly convex function $\varphi : \mathrm{PD}(n) \to \mathbb{R}$. Due to the strict convexity of φ, the function $\varphi(P)$ lies above its tangents $\varphi(Q) + \langle \nabla\varphi(Q), P - Q \rangle$. Hence the non-negativity of the Bregman divergence $D(P, Q)$ is guaranteed.

In this paper, we focus on the Bregman divergence induced from potential function [17]. Let $V : \mathbb{R}_+ \to \mathbb{R}$ be a strictly convex, decreasing, and third order continuously differentiable function. For the derivative V', the inequality $V' < 0$ holds from the assumption. Indeed, the assumption leads to $V' \leq 0$ and $V'' \geq 0$, and if $V'(z_0) = 0$ holds for some $z_0 \in \mathbb{R}_+$, then $V'(z) = 0$ holds for all $z \geq z_0$. Hence V is affine function for $z \geq z_0$. This contradicts the strict convexity of V. We define the functions $\nu_V : \mathbb{R}_+ \to \mathbb{R}$ and $\beta_V : \mathbb{R}_+ \to \mathbb{R}$ such that

$$\nu_V(z) = -zV'(z), \qquad \beta_V(z) = \frac{z\nu_V'(z)}{\nu_V(z)}.$$

The subscript V of ν_V and β_V will be dropped if there is no confusion.

Definition 1 (potential function)**.** *Let $V : \mathbb{R}_+ \to \mathbb{R}$ be a function which is strictly convex, decreasing, and third order continuously differentiable. Suppose that the functions ν and β defined from V satisfy the following*

conditions:

$$\nu(z) > 0, \tag{7}$$

$$\beta(z) < \frac{1}{n} \tag{8}$$

for all $z > 0$ *and*

$$\lim_{z \to +0} \frac{z}{\nu(z)^{n-1}} = 0. \tag{9}$$

Then, V *is called potential function or potential for short. For* $P \in \mathrm{PD}(n)$*, the function* $V(\det P)$ *is also referred to as potential on* $\mathrm{PD}(n)$*.*

As shown in [17], the function $V(\det P)$ is strictly convex in $P \in \mathrm{PD}(n)$ if and only if V satisfies (7) and (8). The condition (9) guarantees the existence of Hessian update formula, which is discussed in Section 3.

Given a potential function V, the Bregman divergence defined from the potential function $\varphi(P) = V(\det P)$ in (6) is denotes as $D_V(P, Q)$, and referred to as V*-Bregman divergence*. The V-Bregman divergence has the form of

$$D_V(P, Q) = V(\det P) - V(\det Q) + \nu(\det Q)\langle Q^{-1}, P\rangle - n\nu(\det Q).$$

Indeed, substituting

$$(\nabla V(\det Q))_{ij} = \frac{dV(\det Q)}{dQ_{ij}} = V'(\det Q)\frac{d\det Q}{dQ_{ij}} = -\nu(\det Q)(Q^{-1})_{ij},$$

into (6), we obtain the expression of $D_V(P, Q)$. Below we show some examples of V-Bregman divergence.

Example 1. *For the negative logarithmic function* $V(z) = -\log(z)$*, we have* $\nu(z) = 1$*. Then* V*-divergence is equal to KL-divergence,*

$$D_V(P, Q) = \mathrm{KL}(P, Q) = \langle P, Q^{-1}\rangle - \log\det(PQ^{-1}) - n.$$

Note that $\mathrm{KL}(P, Q) = \mathrm{KL}(Q^{-1}, P^{-1})$ *holds. Hence,* $\mathrm{KL}(P, Q)$ *is convex in both* P *and* Q^{-1}*.*

Example 2. *For the power potential* $V(z) = (1 - z^\gamma)/\gamma$ *with* $\gamma < 1/n$*, we have* $\nu(z) = z^\gamma$ *and* $\beta(z) = \gamma$*. Then, we obtain*

$$D_V(P, Q) = (\det Q)^\gamma \left\{ \langle P, Q^{-1}\rangle + \frac{1 - (\det PQ^{-1})^\gamma}{\gamma} - n \right\}.$$

The KL-divergence is recovered by taking the limit of $\gamma \to 0$*.*

Example 3. *For $0 \leq a < b$, let $V(z)$ be $V(z) = a\log(az+1) - b\log(z)$. Then $V(z)$ is a convex and decreasing function, and we obtain*

$$\nu(z) = b - a + \frac{a}{az+1} > 0, \qquad \beta(z) = \frac{-a^2 z}{(az+1)(a(b-a)z+b)} \leq 0$$

for $z > 0$. The negative-log potential is derived by setting $a = 0$, $b = 1$. This potential satisfies the inequality $0 < b - a \leq \nu(z) \leq b$. The bounding condition of ν will be assumed in the convergence analysis of Section 4.

We apply V-Bregman divergences to extend quasi-Newton update formula.

3 Extended quasi-Newton update formula

To extend the standard quasi-Newton methods, we consider the optimization problem of the V-Bregman divergence instead of the KL-divergence. Let us define the V-BFGS formula as the optimal solution of the problem,

$$(V\text{-BFGS}) \qquad \min_{B \in \mathrm{PD}(n)} D_V(B, B_k), \quad \text{subject to} \quad Bs_k = y_k. \tag{10}$$

Next we define V-DFP update formula which is an extension of the standard DFP formula (2). Note that KL-divergence satisfies $\mathrm{KL}(P,Q) = \mathrm{KL}(Q^{-1}, P^{-1})$.

Then, the optimization problem associated with the DFP update formula (4) can be extended to the problem,

$$(V\text{-DFP}) \qquad \min_{B \in \mathrm{PD}(n)} D_V(B^{-1}, B_k^{-1}), \quad \text{subject to} \quad Bs_k = y_k. \tag{11}$$

The problem (11) is convex in B^{-1}, since the objective function $D_V(B^{-1}, B_k^{-1})$ is convex in B^{-1} and the constraint $s_k = B^{-1}y_k$ is affine in B^{-1}. Mainly we consider the V-BFGS update formula. The argument on the V-DFP update is almost the same.

Theorem 1. *Let $B_k \in \mathrm{PD}(n)$, and suppose $s_k^\top y_k > 0$. Then the problem (10) has the unique optimal solution $B_{k+1} \in \mathrm{PD}(n)$ satisfying*

$$B_{k+1} = \frac{\nu(\det B_{k+1})}{\nu(\det B_k)} B^{BFGS}[B_k; s_k, y_k] + \left(1 - \frac{\nu(\det B_{k+1})}{\nu(\det B_k)}\right) \frac{y_k y_k^\top}{s_k^\top y_k}. \tag{12}$$

The proof is found in Appendix A.

Note that the V-BFGS update formula is represented by the affine sum of $B^{BFGS}[B_k; s_k, y_k]$ and $y_k y_k^\top / s_k^\top y_k$. This form is equivalent to the self-scaling quasi-Newton update [18, 16] defined as

$$B_{k+1} = \theta_k B^{BFGS}[B_k; s_k, y_k] + (1 - \theta_k)\frac{y_k y_k^\top}{s_k^\top y_k}, \tag{13}$$

where θ_k is a positive real number. In the V-BFGS update formula, the coefficient θ_k is determined from the function ν. The inverse of the matrix (13) is given by

$$B_{k+1}^{-1} = \frac{1}{\theta_k}(B^{BFGS}[B_k; s_k, y_k])^{-1} + \left(1 - \frac{1}{\theta_k}\right)\frac{s_k s_k^\top}{s_k^\top y_k}. \tag{14}$$

As the result, for any $\theta_k > 0$, the matrix B_{k+1} in (13) is positive definite. Indeed, for $0 < \theta_k \leq 1$ the expression (13) guarantees the positive definiteness of B_{k+1}, and for $1 < \theta_k$, the expression (14) implies $B_{k+1} \in \mathrm{PD}(n)$. Therefore B_{k+1} in (12) is also positive definite matrix, since any potential V satisfies $\nu_V > 0$.

In the self-scaling update formula in (13), the choice

$$\theta_k = \frac{s_k^\top y_k}{s_k^\top B_k s_k} \tag{15}$$

is often recommended. As analyzed in [16], however, the self-scaling method with inexact line search for the step length tends to lead the relative inefficiency compared to the standard BFGS method. Following Example 4 below, we prove that the self-scaling method with the scaling parameter (15) is not derived from the V-Bregman divergence.

We present a practical way of computing the Hessian approximation (12). In Eq (12), the optimal solution B_{k+1} appears in both sides, that is, we have only the implicit expression of B_{k+1}. The numerical computation is, however, efficiently performed as well as the standard BFGS update. To compute the update formula B_{k+1}, first we compute $\det B_{k+1}$. The determinant of both sides of (12) leads to

$$\det B_{k+1} = \frac{\det(B^{BFGS}[B_k; s_k, y_k])}{\nu(\det B_k)^{n-1}} \cdot \nu(\det B_{k+1})^{n-1}. \tag{16}$$

Hence, by solving the nonlinear equation

$$z = \frac{\det(B^{BFGS}[B_k; s_k, y_k])}{\nu(\det B_k)^{n-1}} \cdot \nu(z)^{n-1}, \qquad z > 0$$

we can find $\det B_{k+1}$. As shown in the proof of Theorem 1, the function $z/\nu(z)^{n-1}$ is monotone increasing. Hence the Newton method is available to find the root of the above equation efficiently. Once we obtain the value of $\det B_{k+1}$, we can compute the Hessian approximation B_{k+1} by substituting $\det B_{k+1}$ into Eq (12). Figure 3 shows the update algorithm of the V-BFGS formula which exploits the Cholesky decomposition of the approximate Hessian matrix. By maintaining the Cholesky decomposition, we can easily compute the the determinant and the search direction. In the algorithm of Figure 3, we require the Wolfe condition [15, Section 3.1] for the step length α_k. As shown in Section 4, the Wolfe condition is useful to establish the convergence property of the optimization algorithm.

In the same way as the proof of Theorem 1, we obtain the V-DFP update formula defined from (11) such that

$$B_{k+1} = \frac{\nu((\det B_k)^{-1})}{\nu((\det B_{k+1})^{-1})} B^{DFP}[B_k; s_k, y_k] + \left(1 - \frac{\nu((\det B_k)^{-1})}{\nu((\det B_{k+1})^{-1})}\right) \frac{y_k y_k^\top}{s_k^\top y_k}. \tag{17}$$

It is straightforward to unify the V-BFGS method and the V-DFP method in the same way as the standard Broyden family [4]. Let $B^{\mathrm{BFGS}}_{V_1,k+1}$ be the Hessian approximation given by the V-BFGS update formula with the potential $V = V_1$, and $B^{\mathrm{DFP}}_{V_2,k+1}$ be the Hessian approximation given by the V-DFP update formula with the potential $V = V_2$. Then the update formula of the (V_1, V_2)-Broyden family is defined by

$$B_{k+1} \;=\; \vartheta\, B^{(V_1)}_{\mathrm{BFGS},k+1} + (1-\vartheta)\, B^{(V_2)}_{\mathrm{DFP},k+1}, \tag{18}$$

for $\vartheta \in [0,1]$. The (V_1, V_2)-Broyden family is obtained by a convex-full of $B^{BFGS}[B_k; s_k, y_k]$, $B^{DFP}[B_k; s_k, y_k]$ and $y_k y_k^\top / s_k^\top y_k$. The standard Broyden family is recovered by setting $V_1(z) = V_2(z) = -\log z$.

Example 4. *We show the V-BFGS formula derived from the power potential. Let $V(z)$ be the power potential $V(z) = (1 - z^\gamma)/\gamma$ with $\gamma < 1/n$. As shown in Example 2, we have $\nu(z) = z^\gamma$. Due to the equality*

$$\det(B^{BFGS}[B_k; s_k, y_k]) = \det(B_k) \frac{s_k^\top y_k}{s_k^\top B_k s_k}$$

and Eq. (16)*, for the power potential we have*

$$\frac{\nu(\det B_{k+1})}{\nu(\det B_k)} = \left(\frac{s_k^\top y_k}{s_k^\top B_k s_k}\right)^\rho, \qquad \rho = \frac{\gamma}{1 - (n-1)\gamma}.$$

V-BFGS update:

Initialization: The function $\nu(z)$ denotes $-V'(z)z$. Let $B_0 \in \mathrm{PD}(n)$ be a matrix which is an initial approximation of the Hessian matrix, and $L_0 L_0^\top = B_0$ be the Cholesky decomposition of B_0. Let $x_0 \in \mathbb{R}^n$ be an initial point, and set $k = 0$.

Repeat: If stopping criterion is satisfied, go to Output.

1. Let $x_{k+1} = x_k - \alpha_k B_k^{-1} \nabla f(x_k)$, where $\alpha_k \geq 0$ is a step length satisfying the Wolfe condition [15, Section 3.1]. The Cholesky decomposition $B_k = L_k L_k^\top$ is available to compute $B_k^{-1} \nabla f(x_k)$.
2. Set $s_k = x_{k+1} - x_k$ and $y_k = \nabla f(x_{k+1}) - \nabla f(x_k)$.
3. Update L_k to $\bar{L}$ which is the Cholesky decomposition of $B^{BFGS}[B_k; s_k, y_k]$, that is,

$$\bar{L}\bar{L}^\top = B^{BFGS}[B_k; s_k, y_k] = B^{BFGS}[L_k L_k^\top; s_k, y_k].$$

   The Cholesky decomposition with rank-one update is available.
4. Compute

$$C = \frac{(\det \bar{L})^2}{\nu((\det L_k)^2)^{n-1}}$$

   and find the root of the equation

$$C \cdot \nu(z)^{n-1} = z, \qquad z > 0.$$

   Let the solution be z^*.
5. Compute the Cholesky decomposition L_{k+1} such that

$$L_{k+1} L_{k+1}^\top = \frac{\nu(z^*)}{\nu((\det L_k)^2)} \bar{L}\bar{L}^\top + \left(1 - \frac{\nu(z^*)}{\nu((\det L_k)^2)}\right) \frac{y_k y_k^\top}{s_k^\top y_k}.$$

6. $k \leftarrow k + 1$.

Output: Local optimal solution x_k.

Figure 2: Pseudo code of V-BFGS method. The Cholesky decomposition with rank-one update is useful in the algorithm.

Then the V-BFGS update formula is given as

$$B_{k+1} = \left(\frac{s_k^\top y_k}{s_k^\top B_k s_k}\right)^\rho B^{BFGS}[B_k; s_k, y_k] + \left(1 - \left(\frac{s_k^\top y_k}{s_k^\top B_k s_k}\right)^\rho\right)\frac{y_k y_k^\top}{s_k^\top y_k}.$$

For γ such that $\gamma < 1/n$, we have $-1/(n-1) < \rho < 1$. Remember that the standard self-scaling update formula corresponds to the above update with $\rho = 1$. Therefore, the standard self-scaling update formula is not derived from the power potential. Indeed, the power potential with $\rho = 1$ or equivalently $\gamma = 1/n$ is a convex function but not a strictly convex function.

In terms of the self-scaling update formula, we show the following proposition.

Proposition 2. *There does not exist the potential function such that in Eq.* (12) *the equality*

$$\frac{\nu(\det B_{k+1})}{\nu(\det B_k)} = \frac{s_k^\top y_k}{s_k^\top B_k s_k} \tag{19}$$

holds for any $B_k \in \mathrm{PD}(n)$ and any $s_k, y_k \in \mathbb{R}^n$ satisfying $s_k^\top y_k > 0$.

Proof. We have two equalities,

$$\det(B^{BFGS}[B_k; s_k, y_k]) = \det(B_k)\frac{s_k^\top y_k}{s_k^\top B_k s_k},$$
$$\det B_{k+1} = \frac{\det(B^{BFGS}[B_k; s_k, y_k])}{\nu(\det B_k)^{n-1}}\nu(\det B_{k+1})^{n-1}.$$

Hence, we have

$$\left(\frac{\nu(\det B_{k+1})}{\nu(\det B_k)}\right)^{n-1} = \frac{\det B_{k+1}}{\det B_k}\cdot\frac{s_k^\top B_k s_k}{s_k^\top y_k}$$

Suppose that there exists a potential function satisfying (19). Then we have

$$\left(\frac{s_k^\top y_k}{s_k^\top B_k s_k}\right)^{n-1} = \frac{\det B_{k+1}}{\det B_k}\cdot\frac{s_k^\top B_k s_k}{s_k^\top y_k},$$

and hence the equality

$$\det B_{k+1} = \det(B_k)\cdot\left(\frac{s_k^\top y_k}{s_k^\top B_k s_k}\right)^n$$

holds. Substituting the above formula into (19), we have

$$\nu\left(\det(B_k)\left(\frac{s_k^\top y_k}{s_k^\top B_k s_k}\right)^n\right) = \nu(\det B_k)\frac{s_k^\top y_k}{s_k^\top B_k s_k}.$$

Let B_k be a positive definite matrix such that $\det B_k = 1$, and z be $z = \left(\frac{s_k^\top y_k}{s_k^\top B_k s_k}\right)^n$. Then we have $\nu(z) = \nu(1)z^{1/n}$ for $z > 0$. The corresponding β_V is given as $\beta_V(z) = 1/n$, and this does not satisfy the definition of the potential function. □

4 Convergence Analysis

We consider the convergence property of the V-BFGS method. Some standard assumptions about the objective function f are stated below. See Section 6.4 of [15] for details.

Assumption 1. *1. The objective function f is twice continuously differentiable.*

2. Let $\nabla^2 f(x)$ be the Hessian matrix of f at x. For the starting point x_0, the level set $\mathcal{L} = \{x \in \mathbb{R}^n \mid f(x) \leq f(x_0)\}$ is convex, and there exist positive constants m and M such that

$$m\|z\|^2 \leq z^\top \nabla^2 f(x) z \leq M\|z\|^2 \tag{20}$$

holds for all $z \in \mathbb{R}^n$ and $x \in \mathcal{L}$.

The following theorem implies that the sequence $\{x_k\}$ generated by the V-BFGS update formula converges to the local minimizer of f if the function ν_V of a potential V satisfies the bounding condition.

Theorem 3. *Let $B_0 \in \mathrm{PD}(n)$ be an initial matrix and $x_0 \in \mathbb{R}^n$ be a starting point which meets Assumption 1. Suppose that there exist positive constants $L_1, L_2 > 0$ such that $L_1 \leq \nu \leq L_2$. Then the sequence $\{x_k\}$ generated by the V-BFGS update converges to the minimizer x^* of f.*

Lemma 4 (Eq. 6.12 in [15])**.** *Let $\bar{G}$ be the averaged Hessian*

$$\bar{G} = \int_0^1 \nabla^2 f(x_k + \tau s) d\tau, \quad s = x_{k+1} - x_k \in \mathbb{R}^n,$$

then the property $y = \bar{G}s$ follows from Taylor's theorem, where $y = \nabla f(x_{k+1}) - \nabla f(x_k)$.

Using Lemma 4, we prove Theorem 3 in a manner similar to Section 8.4 in [15].

Proof of Theorem 3. Let $B_k, k = 0, 1, 2, \ldots$ be the sequence of approximate Hessian matrices generated by the V-BFGS update formula. We define $\bar{B}_{k+1}$ and $\bar{B}_k$ by $\bar{B}_{k+1} = \frac{1}{\nu(\det B_{k+1})} B_{k+1}$ and $\bar{B}_k = \frac{1}{\nu(\det B_k)} B_k$, respectively. Then the update formula shown in Theorem 1 is represented as

$$\bar{B}_{k+1} = \bar{B}_k - \frac{\bar{B}_k s_k s_k^\top \bar{B}_k}{s_k^\top \bar{B}_k s_k} + \frac{1}{\nu(\det B_{k+1})} \frac{y_k y_k^\top}{s_k^\top y_k}. \tag{21}$$

We compute

$$\psi(\bar{B}_{k+1}) = \mathrm{tr}(\bar{B}_{k+1}) - \log\det \bar{B}_{k+1}.$$

The inequality (20) yields

$$\frac{s_k^\top y_k}{\|s_k\|^2} = \frac{s_k^\top \bar{G} s_k}{\|s_k\|^2} \geq m, \tag{22}$$

$$\frac{\|y_k\|^2}{s_k^\top y_k} = \frac{s_k^\top \bar{G}^2 s_k}{s_k^\top \bar{G} s_k} \leq M. \tag{23}$$

We now define

$$\cos\theta_k = \frac{s_k^\top \bar{B}_k s_k}{\|s_k\| \|\bar{B}_k s_k\|}, \qquad q_k = \frac{s_k^\top \bar{B}_k s_k}{\|s_k\|^2}.$$

Then the trace of $\bar{B}_{k+1}$ is bounded above. Indeed, the inequality

$$\mathrm{tr}(\bar{B}_{k+1}) = \mathrm{tr}(\bar{B}_k) - \frac{\|\bar{B}_k s_k\|^2}{s_k^\top \bar{B}_k s_k} + \frac{\|y_k\|^2}{\nu(\det B_{k+1}) s_k^\top y_k} \leq \mathrm{tr}(\bar{B}_k) - \frac{q_k}{\cos^2\theta_k} + \frac{M}{\nu(\det B_{k+1})},$$

holds, where (23) is used. Using the formula $\det(I + xy^\top + uv^\top) = (1 + x^\top y)(1 + u^\top v^\top) - (x^\top v)(y^\top u)$ for $\bar{B}_{k+1}$, we obtain a lower bound of the determinant $\det(\bar{B}_{k+1})$ such that

$$\det(\bar{B}_{k+1}) = \det(\bar{B}_k) \frac{1}{\nu(\det B_{k+1})} \frac{\|s_k\|^2}{s_k^\top \bar{B}_k s_k} \frac{s_k^\top y_k}{\|s_k\|^2} \geq \det(\bar{B}_k) \frac{m}{q_k \nu(\det B_{k+1})}.$$

These inequalities present an upper bound of $\psi(\bar{B}_{k+1})$,

$$\begin{aligned}
\psi(\bar{B}_{k+1}) &\leq \psi(\bar{B}_k) + \left(\frac{M}{\nu(\det B_{k+1})} - \log \frac{m}{\nu(\det B_{k+1})} - 1 \right) \\
&\quad + \left(1 - \frac{q_k}{\cos^2\theta_k} + \log \frac{q_k}{\cos^2\theta_k} \right) + \log\cos^2\theta_k \\
&\leq \psi(\bar{B}_k) + \left(\frac{M}{L_1} - \log \frac{m}{L_2} - 1 \right) + \log\cos^2\theta_k.
\end{aligned}$$

The second inequality is derived from

$$1 - \frac{q_k}{\cos^2\theta_k} + \log\frac{q_k}{\cos^2\theta_k} \leq 0.$$

As the result we obtain

$$0 < \psi(\bar{B}_{k+1}) \leq \psi(\bar{B}_0) + c(k+1) + \sum_{j=1}^{k} \log\cos^2\theta_j,$$

where c is a positive constant such that $c > \frac{M}{L_1} - \log\frac{m}{L_2} - 1$. Let us then proceed by contradiction and assume that $\cos\theta_j \to 0$. Then there exists $k_1 > 0$ such that for all $j > k_1$, we have

$$\log\cos^2\theta_j < -2c.$$

Thus the following inequality holds for all $k > k_1$:

$$\begin{aligned} 0 &< \psi(\bar{B}_0) + c(k+1) + \sum_{j=1}^{k_1} \log\cos^2\theta_j + (k-k_1)(-2c) \\ &= \psi(\bar{B}_0) + \sum_{j=1}^{k_1} \log\cos^2\theta_j + c(2k_1+1) - 2ck. \end{aligned}$$

The right-hand-side is negative for large k, giving a contradiction. Therefore there exists a subsequence satisfying $\cos\theta_{j_k} \geq \delta > 0$. By Zoutendijk's result[1] with the Wolfe condition, this limit implies that $\liminf_{k\to\infty} \|\nabla f(x_k)\| = 0$. The convexity of f on $\mathcal{L}$ guarantees that x_k converges to the local optimal solution. □

The potential defined in Example 3 meets the condition of Theorem 3, while the power potential $V(z) = (1 - z^\gamma)/\gamma$ with $\nu(z) = z^\gamma$ does not satisfy the condition.

5 Robustness against Inexact Line Search

The robustness against numerical errors such as the round-off error is an important feature in numerical computation. In this section we study the robustness of quasi-Newton update against numerical errors involved in the line search. Mainly there are two types of quasi-Newton updates: one is the

[1] Under some condition, $\sum_{j\geq 0} \cos^2\theta_j \|\nabla f(x_j)\|^2 < \infty$ holds. See Theorem 3.2 in [15]

update formula for approximate Hessian matrix; and the other is the update for approximate *inverse* Hessian matrix. In the approximate inverse Hessian update, the matrix $H_k = B_k^{-1}$ is directly update to $H_{k+1} = B_{k+1}^{-1}$ under the secant condition $H_{k+1}y_k = s_k$. We study four kinds of update formulae, that is, V-BFGS/V-DFP method for the Hessian approximation/the inverse Hessian approximation.

Let us consider the Hessian approximation formula. Under the exact line search, the matrix B_k is updated to B_{k+1} which is the minimum solution of $D_V(B, B_k)$ or $D_V(B^{-1}, B_k^{-1})$ subject to $Bs_k = y_k$. Let

$$x_{k+1} = x_k - \alpha_k B_k^{-1}\nabla f(x_k) = x_k + s_k$$

be the point computed by the exact line search. When the line search is inexact, the step length α_k will be slightly perturbed and then s_k will be changed to $(1+\varepsilon)s_k$ where ε is an infinitesimal. The vector y_k will also change to $\widetilde{y}_k$ defined by

$$\widetilde{y}_k = \nabla f(x_k + (1+\varepsilon)s_k) - \nabla f(x_k) = y_k + \varepsilon\nabla^2 f(x_{k+1})s_k + O(\varepsilon^2).$$

Then the constraint for the Hessian update becomes $(1+\varepsilon)Bs_k = \widetilde{y}_k$.

We study the relation between the perturbation of s_k and the Hessian approximation B_{k+1} or the inverse Hessian approximation H_{k+1}. Based on the above argument, we consider the optimization problem defined by

$$(V\text{-BFGS-B}) \qquad \min_{B\in \mathrm{PD}(n)} D_V(B, B_k) \quad \text{subject to} \quad (1+\varepsilon)Bs = y + \varepsilon\bar{y}, \tag{24}$$

$$(V\text{-DFP-B}) \qquad \min_{B\in \mathrm{PD}(n)} D_V(B^{-1}, B_k^{-1}) \quad \text{subject to} \quad (1+\varepsilon)Bs = y + \varepsilon\bar{y} \tag{25}$$

for a fixed matrix $B_k \in \mathrm{PD}(n)$ and fixed vectors $s, y, \bar{y} \in \mathbb{R}^n$, where the subscript k for the vectors is dropped for simplicity. In the same way, the update formula for the inverse Hessian under the inexact line search is defined as the optimal solution of the following problem,

$$(V\text{-BFGS-H}) \qquad \min_{H\in \mathrm{PD}(n)} D_V(H^{-1}, H_k^{-1}) \quad \text{subject to} \quad H(y + \varepsilon\bar{y}) = (1+\varepsilon)s, \tag{26}$$

$$(V\text{-DFP-H}) \qquad \min_{H\in \mathrm{PD}(n)} D_V(H, H_k) \quad \text{subject to} \quad H(y + \varepsilon\bar{y}) = (1+\varepsilon)s, \tag{27}$$

for fixed $H_k \in \mathrm{PD}(n)$, $s, y, \bar{y} \in \mathbb{R}^n$. The update formula given by V-BFGS-H/V-DFP-H directly provides the inverse matrix of B_{k+1} computed by V-BFGS-B/V-DFP-B, respectively. Theorem 1 guarantees that there exists the unique optimal solution as long as $s^\top(y + \varepsilon\bar{y}) > 0$ holds. Though Theorem 1 deals with only V-BFGS-B formula, we can prove the existence and the uniqueness of optimal solution for the other problems in the same manner.

In order to study the robustness of update formulae, we borrow the concepts such that the influence function or the gross error sensitivity from the study of robust statistics [11]. Below the V-BFGS-B update formula is considered as an example. Let $B(\varepsilon)$ be the optimal solution of V-BFGS-B in (24). Then the *influence function* of $B(\varepsilon)$ is defined as the derivative of $B(\varepsilon)$ at $\varepsilon = 0$, that is,

$$\dot{B}(0) = \lim_{\varepsilon \to 0} \frac{B(\varepsilon) - B(0)}{\varepsilon}.$$

Later we prove the differentiability of $B(\varepsilon)$. From the definition of the influence function, the optimal solution $B(\varepsilon)$ is asymptotically equal to $B(0) + \varepsilon\dot{B}(0)$. This implies that the inexact line search has a large impact on the computation of Hessian approximation, when the norm of $\dot{B}(0)$ is large. In the sense of the influence function, the preferable potential is the function V which provides the influence function $\dot{B}(0)$ with a small norm.

For fixed vectors s and y such that $s^\top y > 0$, the influence function $\dot{B}(0)$ depends on the matrix B_k and the vector $\bar{y}$. We consider the worst-case evaluation of the influence function in terms of B_k and $\bar{y}$. The *gross error sensitivity* is defined as the largest norm of the influence function, that is,

$$\text{gross error sensitivity} = \sup\left\{ \|\dot{B}(0)\|_F \mid B_k \in \mathcal{B} \subset \mathrm{PD}(n),\, \bar{y} \in \mathcal{Y} \subset \mathbb{R}^n \right\},$$

where $\mathcal{B} \subset \mathrm{PD}(n)$ and $\mathcal{Y} \subset \mathbb{R}^n$ are appropriate subsets. In many case, the gross error sensitivity becomes infinity if $\mathcal{B}$ or $\mathcal{Y}$ is unbounded. Our concern is to find the potential function V which leads finite gross error sensitivity under some reasonable setup.

The influence function and the gross error sensitivity have been studied in robust statistics [11]. We use these statistical techniques to analyze the stability of numerical computation. In the literature of statistics, the "statistical model" $\{B \in \mathrm{PD}(n) \mid Bs_k = y_k\}$ or $\{H \in \mathrm{PD}(n) \mid Hy_k = s_k\}$ is fixed, and the "observed data" B_k or H_k is contaminated such that $B_k + \varepsilon\dot{B}(0) + O(\varepsilon^2)$, while in the present analysis, the matrix $B_k = H_k^{-1}$ is fixed and the model corresponding to the secant condition is perturbed.

Table 1: Gross error sensitivity of V-BFGS formula and V-DFP formula for the Hessian approximation and the inverse Hessian approximation. Only the standard BFGS for the Hessian approximation has finite gross error sensitivity.

	V-BFGS	V-DFP
Hessian approx.	finite only for BFGS	∞
inverse Hessian approx.	∞	∞

The potential function minimizing the gross error sensitivity will be preferable for robust computation. Below we prove that the standard BFGS update for the Hessian approximation is the more robust than the other update formulae. This result meets the empirical observations [5, 15]. Moreover, only the standard BFGS update for the Hessian approximation has finite gross error sensitivity. Theoretical results are summarized in Table 1.

In the following, the gross error sensitivity with $\mathcal{B} = \mathrm{PD}(n)$ and a bounded subset $\mathcal{Y}$ is considered. Note that the boundedness of $\mathcal{Y}$ follows the assumption that $\|\nabla^2 f\|_F$ is bounded above over $\mathbb{R}^n$. First, we note that the influence function and the gross error sensitivity make sense for minimization of non-quadratic functions.

Lemma 5. *Suppose that the objective function $f(x)$ is a convex quadratic function. Then, the influence function and the gross error sensitivity are equal to zero.*

Lemma 5 is clear, since for the quadratic objective function the secant condition $Bs = y$ is changed to $B(1+\varepsilon)s = (1+\epsilon)y$ under the inexact line search. That is, the secant condition is kept unchanged, and thus $B(\varepsilon) = B(0)$ holds.

We prove that generally the influence function is well-defined.

Theorem 6. *Suppose that $s^\top y > 0$ holds for vectors s and y in the problems* (24), (25), (26) *and* (27). *Then, for small ε, the optimal solutions of V-BFGS-B, V-DFP-B, V-BFGS-H and V-DFP-H are all uniquely determined. The optimal solutions are second-order continuously differentiable with respect to ε in the vicinity of $\varepsilon = 0$.*

Proof is deferred to Appendix B.

The gross error sensitivity of each update formula is computed in the following theorems. Proofs are deferred to Appendix C.

Theorem 7 (gross error sensitivity of V-BFGS-B)**.** *Suppose $n \geq 3$. Let s and y be fixed vectors such that $s^\top y > 0$ and $\mathcal{Y}$ be a bounded subset in $\mathbb{R}^n$. For small ε, let $B(\varepsilon)$ be the optimal solution of V-BFGS-B in* (24)*. Then, the optimal potential function of the problem*

$$\min_V \max_{B_k, \bar{y}} \|\dot{B}(0)\|_F \quad \text{subject to} \quad B_k \in \mathrm{PD}(n), \quad \bar{y} \in \mathcal{Y} \tag{28}$$

is given as $V(z) = -\log(z)$ up to a constant factor. In the above min-max problem, the function V is sought from among all potentials.

Theorem 8 (gross error sensitivity of V-DFP-B)**.** *Suppose $n \geq 3$. Let s and y be fixed vectors such that $s^\top y > 0$ and $\mathcal{Y}$ be a bounded subset in $\mathbb{R}^n$. Suppose that there exists an open subset included in $\mathcal{Y}$. Let $B(\varepsilon)$ be the optimal solution of V-DFP-B in* (25)*. Then for any potential V, the equality*

$$\sup\{\|\dot{B}(0)\|_F \mid B_k \in \mathrm{PD}(n),\, \bar{y} \in \mathcal{Y}\} = \infty$$

holds.

Theorem 9 (gross error sensitivity of V-BFGS-H)**.** *Suppose $n \geq 4$. Let s and y be fixed vectors such that $s^\top y > 0$ and $\mathcal{Y}$ be a bounded subset in $\mathbb{R}^n$. Suppose that there exists an open subset included in $\mathcal{Y}$. Let $H(\varepsilon)$ be the optimal solution of V-BFGS-H in* (26)*. Then, for any potential V, the equality*

$$\sup\{\|\dot{H}(0)\|_F \mid H_k \in \mathrm{PD}(n),\, \bar{y} \in \mathcal{Y}\} = \infty$$

holds.

Theorem 10 (gross error sensitivity of V-DFP-H)**.** *Suppose $n \geq 3$. Let s and y be fixed vectors such that $s^\top y > 0$ and $\mathcal{Y}$ be a bounded subset in $\mathbb{R}^n$. Let $H(\varepsilon)$ be the optimal solution of V-DFP-H in* (27)*. Then, for any potential V, the equality*

$$\sup\{\|\dot{H}(0)\|_F \mid H_k \in \mathrm{PD}(n),\, \bar{y} \in \mathcal{Y}\} = \infty$$

holds.

It is well-known that there is the dual relation between the BFGS formula and the DFP formula. Indeed, the V-DFP update for the inverse Hessian approximation is derived from the V-BFGS update formula for the Hessian approximation by replacing B_k, s_k, y_k with H_k, y_k, s_k. For the robustness

against inexact line search, however, the dual relation is violated as shown in Table 1. In this problem, we focus on the perturbation of the vector s_k rather than that of y_k. This is the reason why the dual relation is violated. Powell has shown a critical difference between BFGS and DFP for quadratic convex objective functions [19] by considering the behaviour of eigenvalues of approximate Hessian matrix. In the present paper, we exploited the gross error sensitivity which is meaningful for non-quadratic objective functions as shown in Lemma 5. Our approach also provides a critical difference between BFGS and DFP methods.

In Section 3, we introduced the (V_1, V_2)-Broyden family defined by (18). It is straightforward to prove that only the standard BFGS has finite gross error sensitivity among the (V_1, V_2)-Broyden family with a fixed mixing parameter $\vartheta \in [0, 1]$.

6 Numerical Studies

We demonstrate numerical experiments on robustness of quasi-Newton update formulae such as V-BFGS-B, V-DFP-B, V-BFGS-H, and V-DFP-H proposed in Section 5. Especially, the update formula derived from power potential in Example 2 is examined.

In the first numerical study, we consider numerical stability of update formulae. Let $B(\varepsilon)$ be the optimal solution of V-BFGS-B (24) or V-DFP-B (25), and $H(\varepsilon)$ be the optimal solution of V-BFGS-H (26) or V-DFP-H (27). For each update formula, we numerically compute the approximate influence function $\|(B(\varepsilon)-B(0))/\varepsilon\|_F$ and $\|(H(\varepsilon)-H(0))/\varepsilon\|_F$ with small ε, where the power potential $V(z) = (1-z^\gamma)/z$ is used to derive the approximate Hessian matrix. Remember that V-BFGS and V-DFP are respectively reduced to the standard BFGS and DFP when γ is equal to zero.

In what follows, we show the setup of numerical studies. Let $\mathrm{diag}(a_1, \ldots, a_n)$ be the n by n diagonal matrix with diagonal elements $a_1, \ldots, a_n$. For V-BFGS-B and V-DFP-B, the matrix B_k is set to one of the following three matrices:

$$B_k = \mathrm{diag}(1, \ldots, n)/(n!)^{1/n}, \quad B_k = \mathrm{diag}(1, \ldots, n), \quad \text{or} \quad B_k = I + n^3 \cdot pp^\top,$$

where in the last one I is the identity matrix and p is a column unit vector defined below. The dimension of the matrix B_k is set to $n = 10, 100, 500$ or 1000. The first matrix $\mathrm{diag}(1, \ldots, n)/(n!)^{1/n}$ has the determinant one, and the other two matrices have a large determinant. Below we show the procedure for generating the vectors s and y and the contaminated vectors

$(1+\varepsilon)s$ and $y+\varepsilon\bar{y}$ for V-BFGS-B and V-DFP-B. In the numerical studies for V-BFGS-H and V-DFP-H, the matrix B_k is replaced with the approximate inverse Hessian H_k.

1. In the case that B_k is $\mathrm{diag}(1,\ldots,n)/(n!)^{1/n}$ or $\mathrm{diag}(1,\ldots,n)$, the vectors s and y are both generated according to the multivariate normal distribution with mean zero and variance-covariance matrix $10 \times I$. If the inner product $s^\top y$ is non-positive, the sign of y is flipped. The intensity of noise involved in the line search is determined by ε, which is generated according to the uniform distribution on the interval $[-0.2, 0.2]$. Then, the vector $\bar{y}$ is also generated according to the multivariate standard normal distribution. If the inequality $(1+\varepsilon)s^\top(y+\varepsilon\bar{y}) > 0$ does not hold, again ε and $\bar{y}$ are generated until the vectors enjoy the positivity condition.

2. In the case that B_k is supposed to have the expression $I + n^3 \cdot pp^\top$, first the vector s is generated according to the multivariate normal distribution with mean zero and variance-covariance matrix $10 \times I$, and y is defined such that $y = s$. The vector p is a unit vector which is orthogonal to y, that is, p is a vector satisfying $p^\top y = 0$ and $\|p\| = 1$, and let B_k be $B_k = I + n^3 \cdot pp^\top$. Then the vector $\bar{y}$ is defined as $\bar{y} = p$. The construction of these vectors is used in the proof of Theorem 9 and Theorem 10.

Hessian or inverse Hessian update formula is applied to B_k or H_k with the randomly generated secant condition. The updated matrix $B(0)$ and $B(\varepsilon)$ are respectively computed under the constraint $Bs = y$ and $B(1+\varepsilon)s = y + \varepsilon\bar{y}$ by using V-BFGS-B and V-DFP-B update formula. In the same way, V-BFGS-H and V-DFP-H are respectively applied to compute $H(0)$ with the constraint $Hy = s$ and $H(\varepsilon)$ with the perturbed secant condition $H(y + \varepsilon\bar{y}) = (1+\varepsilon)s$. The influence function of each update formula is approximated by $\|(B(\varepsilon) - B(0))/\varepsilon\|_F$ or $\|(H(\varepsilon) - H(0))/\varepsilon\|_F$.

Table 2 shows the average of the approximate influence function over 20 runs for each setup. When B_k or H_k is equal to the diagonal matrix $\mathrm{diag}(1,\ldots,n)/(n!)^{1/n}$, we see that the power γ of the power potential does not significantly affect the influence function in both V-BFGS and V-DFP. For the other setups, overall the BFGS method for Hessian matrix, i.e. V-BFGS-B with $\gamma = 0$, has smaller influence function than the other update formulae. The V-DFP-H for inverse Hessian update also has relatively small influence function when H_k is proportional to $\mathrm{diag}(1,\ldots,n)$. For $H_k =$

$I+n^3pp^\top$, however, we find that V-DFP-H is sensitive against noise involved in the line search.

These numerical results meet the theoretical analysis as shown below:

1. Theorem 7 implies that the standard BFGS method is robust against inexact line search.

2. As shown in Example 4, V-BFGS-B update with power potential is close to the standard BFGS update for large n and moderate $\det(B_k)$. That is, the mixing parameter $(s_k^\top y_k/s_k^\top B_k s_k)^\rho$ in Example 4 will be close to one if n is large and $s_k^\top y_k/s_k^\top B_k s_k$ does not depend on the dimension n that much. When B_k has a large determinant which grows with the dimension n, the number of $s_k^\top y_k/s_k^\top B_k s_k$ will severely depend on the dimension n. Hence, the mixing parameter $(s_k^\top y_k/s_k^\top B_k s_k)^\rho$ will not close to one even for large n. Hence, in such case the influence function is affected by the choice of the power γ. The same argument on the relation between influence function and the power γ will hold for the inverse Hessian update, that is, V-BFGS-H and V-DFP-H.

3. For $B_k = I + n^3pp^T$ the result on V-BFGS-B and V-DFP-B is numerically the same. Under this setup, we can theoretically confirm that the influence functions of both update formula are identical to each other. On the other hand, some calculation yields that the influence functions of V-BFGS-H and V-DFP-H are not the same.

The standard BFGS update formula achieves the min-max optimality of the gross error sensitivity. That is, BFGS method may not be necessarily optimal for each setup. In numerical studies, however, BFGS method uniformly provides fairly stable update formula compared to the other methods.

Next, we apply the standard BFGS-B and DFP-B to solve the following two optimization problems: the quadratic convex problem

$$\text{(Problem 1)} \qquad \min_{x\in\mathbb{R}^n} \; f(x) = \frac{1}{2}x^\top Ax - e^\top x,$$

where $e = (1,\dots,1)^\top \in \mathbb{R}^n$ and

$$A = \begin{pmatrix} 2 & -1 & & & \\ -1 & 2 & -1 & & \\ & -1 & 2 & \ddots & \\ & & \ddots & \ddots & -1 \\ & & & -1 & 2 \end{pmatrix} \in \mathbb{R}^{n\times n},$$

and the boundary value problem [8]

$$\text{(Problem 2)} \qquad \min_{x\in\mathbb{R}^n} \; f(x) = \frac{1}{2}x^\top Ax - e^\top x - \frac{1}{(n+1)^2}\sum_{i=1}^{n}(2x_i + \cos x_i),$$

where the vector e and the matrix A are the same as problem 1. The objective function in problem 2 is non-linear and non-convex. The initial point x_0 is randomly generated by n-dimensional normal distribution with mean zero and variance-covariance matrix $10 \times I$. The termination criterion

$$\|\nabla f(x_k)\| \leq n \times 10^{-5} \quad \text{or} \quad k \geq 50000,$$

is employed, which is the same criterion used by Yamashita [20]. Although the second criterion above implies that the method fails to obtain a solution, all trials did not reach the maximum number of iterations. In each problem, the step-length α_k is computed by the matlab command "`fminbnd`" with the option `TolX` $= 10^{-12}$ which denotes the termination tolerance on x. In the same way as the numerical studies on robustness of update formulae, the vector $s_k = x_{k+1} - x_k$ is randomly perturbed such that $\widetilde{s}_k = (1+\varepsilon)s_k$, where ε is a random variable according to the uniform distribution on the interval $[-h, h]$. The number of h varies from 0 to 0.3. Accordingly, the vector $y_k = \nabla f(x_{k+1}) - \nabla f(x_k)$ is also changed to $\widetilde{y}_k = \nabla f(x_k + \widetilde{s}_k) - \nabla f(x_k)$. As the result, for each iteration the secant condition with inexact line search is given as $B\widetilde{s}_k = \widetilde{y}_k$.

The average number of iterations over 20 runs for BFGS and DFP is shown in Table 3. Compared to DFP method, BFGS method requires fewer number of iterations to reach the optimal solution. Moreover, in BFGS update the number of iterations is stable against the number of h. This result implies that BFGS is robust against random noise involved in inexact line search. On the other hand, the behaviour of DFP method is sensitive to contaminated step-length. Indeed, the number of iterations in DFP method rises drastically with the intensity of the noise. For the quadratic convex objective function, the inexact line search does not affect the secant condition. Hence the numerical result will impliy that the goodness of the descent direction $B_k^{-1}\nabla f(x_k)$ in DFP will be easily degraded by inexact line search. These numerical properties in quasi-Newton methods have been empirically well-known [5, 15]. Powell [19] has theoretically studied the progression of eigenvalues in approximate Hessian matrices in order to illustrate the difference between BFGS and DFP.

Through the numerical stduies in this section, we found that the theoretical framework exploiting robust statistics can be a useful tool to investigate the property of quasi-Newton methods.

7 Concluding Remarks

Along the line of the research stared by Fletcher [7], we considered the quasi-Newton update formula based on the Bregman divergence induced from potential functions. The proposed update formulae for the Hessian approximation belong to the class of self-scaling quasi-Newton method. We studied the convergence property. Then, we applied the tools in the robust statistics to analyze the robustness of the Hessian update formulae. As the result, we found that the influence of the inexact line search is bounded only for the standard BFGS formula for the Hessian approximation. Numerical studies support the usefulness of the theoretical framework borrowed from the robust statistics.

It will be an interesting future work to investigate the practical advantage of the self-scaling quasi-Newton methods derived from the V-Bregman divergence. Nocedal and Yuan proved that the self-scaling quasi-Newton method with the popular scaling parameter (15) has some drawbacks [16]. In our framework, the self-scaling quasi-Newton method with the scaling parameter (15) is out of the formulae derived from V-Bregman divergence. More precisely, the function $V(z) = n(1 - z^{1/n})$, which is not potential, formally leads the popular self-scaling quasi-Newton formula. For the corresponding Bregman divergence $D_V(P, Q)$, the equality $D_V(P, cP) = 0$ holds for any $P \in \mathrm{PD}(n)$ and any $c > 0$. This property implies that the scale of the Hessian approximation is not fixed. We think that this property may lead some inefficiency of the self-scaling quasi-Newton method with (15). The self-scaling quasi-Newton method associated with V-Bregman divergence may performs well in practice.

Another research direction is to consider the choice of the potential function V. Under the criterion of the gross error sensitivity, we found that the negative logarithmic function $V(z) = -\log z$ is the optimal choice. The other criterion may lead other optimal potentials. Investigating the relation between the criterion for the update formula and the optimal potential will be beneficial for the design of numerical algorithms.

8 Acknowledgements

The authors are grateful to Dr. Nobuo Yamashita of Kyoto university for helpful comments. T. Kanamori was partially supported by Grant-in-Aid for Young Scientists (20700251).

A Proof of Theorems 1

We prove the following lemma which is useful to show the existence of the optimal solution.

Lemma 11. *Let V be a potential and $\nu = \nu_V$. For any $C > 0$ the equation*

$$C\nu(z)^{n-1} = z, \quad z > 0 \tag{29}$$

has the unique solution.

Proof. We define the function $\zeta(z)$ by $\zeta(z) = \log z - (n-1)\log \nu(z)$, then, the (29) is equivalent to the equation

$$\log C = \zeta(z), \quad z > 0. \tag{30}$$

Since the potential function satisfies $\lim_{z\to+0} z/\nu(z)^{n-1} = 0$ from the definition, we have $\lim_{z\to+0}\zeta(z) = -\infty$. In terms of the derivative of $\zeta(z)$, we have the following inequality

$$\frac{d}{dz}\zeta(z) = \frac{1}{z} - (n-1)\frac{\beta(z)}{z} > \frac{1}{zn} > 0.$$

Thus, $\zeta(z)$ is an increasing function on $\mathbb{R}_+$. Moreover we have

$$\zeta(z) \geq \zeta(1) + \int_1^z \frac{1}{zn} dz = \zeta(1) + \frac{\log z}{n}.$$

The above inequality implies that $\lim_{z\to\infty}\zeta(z) = \infty$. Since $\zeta(z)$ is continuous, the equation (30) has the unique solution. □

Proof of Theorem 1. First, we show the existence of the matrix B_{k+1} satisfying (12). Lemma 11 now shows that there exists a solution $z^* > 0$ for the equation

$$\frac{\det(B^{BFGS}[B_k; s_k, y_k])}{\nu(\det B_k)^{n-1}} \cdot \nu(z)^{n-1} = z, \quad z > 0.$$

By using the solution z^*, we define the matrix $\bar{B}$ such that

$$\bar{B} = \frac{\nu(z^*)}{\nu(\det B_k)} B^{BFGS}[B_k; s_k, y_k] + \big(1 - \frac{\nu(z^*)}{\nu(\det B_k)}\big)\frac{y_k y_k^\top}{s_k^\top y_k},$$

then the determinant of $\bar{B}$ satisfies

$$\det \bar{B} = \frac{\det(B^{BFGS}[B_k])}{\nu(\det B_k)^{n-1}} \cdot \nu(z^*)^{n-1} = z^*,$$

in which the first equality comes from the formula $\det(A+vu^\top) = \det(A)(1+u^\top A^{-1}v)$ and the second one follows the definition of z^*. Hence there exists $B_{k+1} \in \mathrm{PD}(n)$ satisfying (12).

Next, we show that the matrix B_{k+1} in (12) satisfies the optimality condition of (10). According to Güler, et al. [10], the normal vector for the affine subspace

$$\mathcal{M} = \{B \in \mathrm{PD}(n) \mid Bs_k = y_k\}$$

is characterized by the form of

$$s_k\lambda^\top + \lambda s_k^\top \in \mathrm{Sym}(n), \qquad \lambda \in \mathbb{R}^n. \tag{31}$$

In fact for $B_1, B_2 \in \mathcal{M}$ we have

$$\begin{aligned}\langle s_k\lambda^\top + \lambda s_k^\top,\, B_1 - B_2\rangle &= \lambda^\top B_1 s_k + s_k^\top B_1\lambda - \lambda^\top B_2 s_k - s_k^\top B_2\lambda \\ &= \lambda^\top y_k + y_k^\top\lambda - \lambda^\top y_k - y_k^\top\lambda \\ &= 0,\end{aligned}$$

and thus $s_k\lambda^\top + \lambda s_k^\top$ is a normal vector of $\mathcal{M}$. Güler, et al. [10] have shown that the normal vector is restricted to the form of (31).

Suppose $B' \in \mathrm{PD}(n)$ be an optimal solution of (10), then B' satisfies the optimality condition that there exists a vector $\lambda \in \mathbb{R}^n$ such that

$$\begin{aligned}&\nabla_B D_V(B, B_k)\big|_{B=B'} = s_k\lambda^\top + \lambda s_k^\top \\ \iff\ &-\nu(\det(B'))(B')^{-1} + \nu(\det(B_k))B_k^{-1} = s_k\lambda^\top + \lambda s_k^\top,\end{aligned}$$

where $\nabla_B D_V(B, B_k)$ denotes the gradient of $D_V(B, B_k)$ with respect to the variable B. Also, the optimal solution B' should satisfy the constraint $B's_k = y_k$. On the other hand, the matrix B_{k+1} defined by (12) satisfies

$$\begin{aligned}B_{k+1}^{-1} &= \frac{\nu(\det B_k)}{\nu(\det B_{k+1})}(B^{BFGS}[B_k; s_k, y_k])^{-1} + \left(1 - \frac{\nu(\det B_k)}{\nu(\det B_{k+1})}\right)\frac{s_k s_k^\top}{s_k^\top y_k} \\ &= \frac{\nu(\det B_k)}{\nu(\det B_{k+1})}B^{DFP}[B_k^{-1}; y_k, s_k] + \left(1 - \frac{\nu(\det B_k)}{\nu(\det B_{k+1})}\right)\frac{s_k s_k^\top}{s_k^\top y_k} \\ \iff &\begin{cases} -\nu(\det B_{k+1})B_{k+1}^{-1} + \nu(\det B_k)B_k^{-1} = s_k\lambda^\top + \lambda s_k^\top, \\ \lambda = \dfrac{\nu(\det B_k)}{s_k^\top y_k}B_k^{-1}y_k - \dfrac{\nu(\det B_{k+1})}{2s_k^\top y_k}s_k - \dfrac{\nu(\det B_k)y_k^\top B_k^{-1}y_k}{2(s_k^\top y_k)^2}s_k. \end{cases}\end{aligned}$$

The conditions $s_k^\top y_k > 0$ and $B_k \in \mathrm{PD}(n)$ guarantees the existence of the above vector λ. In addition, the direct computation yields that the constraint $B_{k+1}s_k = y_k$ is satisfied. Hence, B_{k+1} satisfies the optimality condition. Since (10) is a strictly convex problem, B_{k+1} is the unique optimal solution. □

B Proofs of Theorems 6

We show that the optimal solution of V-BFGS-B is second order continuously differentiable. The same proof works for the other update formulae.

Proof. We consider the problem (24). Since the inequality $s^\top(y+\varepsilon\bar{y}) > 0$ holds for infinitesimal ε, Theorem 1 guarantees that there exists the unique optimal solution $B(\varepsilon)$ around $\varepsilon = 0$. Let the function $F : \mathbb{R}^{n\times n} \times \mathbb{R} \to \mathbb{R}^{n\times n}$ be

$$F(X,\varepsilon) = \frac{1}{\nu(\det X)}X - \frac{1}{\nu(\det B_k)}B^{BFGS}[B_k; (1+\varepsilon)s, y+\varepsilon\bar{y}] - \left(\frac{1}{\nu(\det X)} - \frac{1}{\nu(\det B_k)}\right)\frac{(y+\varepsilon\bar{y})(y+\varepsilon\bar{y})^\top}{(1+\varepsilon)s^\top(y+\varepsilon\bar{y})},$$

for $X \in \mathbb{R}^{n\times n}$ and $\varepsilon \in \mathbb{R}$. For infinitesimal ε, the equality $F(B(\varepsilon),\varepsilon) = O$ holds, where O is the null matrix. We apply the implicit function theorem to prove the differentiability of $B(\varepsilon)$. Since the potential function is third order continuously differentiable, clearly $F(X,\varepsilon)$ is second order continuously differentiable in a vicinity of $(X,\varepsilon) = (B(0),0)$. For any symmetric matrix $A \in \mathrm{Sym}(n)$, the equality

$$\nabla_X \langle F(X,\,\varepsilon),\, A\rangle\big|_{X=B(0),\varepsilon=0} = \frac{1}{\nu(\det B(0))}A - \frac{1}{\nu(\det B(0))^2}\left\langle B(0) - \frac{yy^\top}{s^\top y},\, A\right\rangle B(0)^{-1}$$

holds, where ∇_X denotes the gradient with respect to the variable X. This implies that the gradient of $F(X,\varepsilon)$ does not vanish at $(X,\varepsilon) = (B(0),0)$. Hence, the implicit function theorem for $F(X,\varepsilon)$ guarantees that $B(\varepsilon)$ is a second order continuously differentiable function with respect to ε in a vicinity of $\varepsilon = 0$. □

C Computations of Gross Error Sensitivity

First, a universal formula for the computation of influence function is proved, and some useful lemmas are prepared. Then, the gross error sensitivity for each update formula is computed in Section C.1, C.2, C.3 and C.4.

Lemma 12. *Let $s, \bar{s}, y$ and $\bar{y}$ be column vectors in $\mathbb{R}^n$ such that $s^\top y > 0$, and B_k be a positive definite matrix. For an infinitesimal ε let $B(\varepsilon)$ be the optimal solution of*

$$\min_{B\in\mathrm{PD}(n)} D_V(B, B_k) \quad \text{subject to } B(s+\varepsilon\bar{s}) = y+\varepsilon\bar{y}, \tag{32}$$

and let $\Delta[B_k; s, \bar{s}, y, \bar{y}]$ *be the influence function* $\dot{B}(0)$*. Then we have*

$$\begin{aligned}
&\dot{B}(0)\\
&= \Delta[B_k; s, \bar{s}, y, \bar{y}]\\
&= \left\{ \frac{s^\top \bar{y} - \bar{s}^\top y}{s^\top y} + \frac{\nu(\det B(0))}{\nu(\det B_k)} \left(\frac{2\bar{s}^\top B_k s \cdot s^\top B_k (B(0))^{-1} B_k s}{(s^\top B_k s)^2} - \frac{2\bar{s}^\top B_k (B(0))^{-1} B_k s}{s^\top B_k s} \right) \right\}\\
&\quad \times \frac{\beta(\det B(0))}{1-(n-1)\beta(\det B(0))} \left[B(0) - \frac{yy^\top}{s^\top y} \right] + \frac{y\bar{y}^\top + \bar{y}y^\top}{s^\top y} - \frac{s^\top \bar{y} + \bar{s}^\top y}{(s^\top y)^2} yy^\top\\
&\quad + \frac{\nu(\det B(0))}{\nu(\det B_k)} \left[\frac{2\bar{s}^\top B_k s}{(s^\top B_k s)^2} B_k s s^\top B_k - \frac{B_k (s\bar{s}^\top + \bar{s}s^\top) B_k}{s^\top B_k s} \right]. \qquad (33)
\end{aligned}$$

The matrix $\Delta[B_k; s, \bar{s}, y, \bar{y}]$ is well-defined, since the inequalities $\nu > 0$ and $1-(n-1)\beta > 0$ hold for any potential function. Note that $\Delta[B_k; s, s, y, y] = O$ holds. This is another proof of Lemma 5.

Proof of Lemma 12. In the same way as the proof of Theorem 1 and Theorem 6, we can prove the existence and the differentiability of $B(\varepsilon)$. Since $B(\varepsilon)$ is second order continuously differentiable around $\varepsilon = 0$, the equality

$$B(\varepsilon) = B(0) + \varepsilon\Delta + O(\varepsilon^2),$$

holds, where $\Delta \in \mathrm{Sym}(n)$. Then we have

$$\begin{aligned}
\det(B(\varepsilon)) &= \det(B(0) + \varepsilon\Delta + O(\varepsilon^2))\\
&= \det(B(0)) + \varepsilon \det(B(0)) \langle \Delta, B(0)^{-1} \rangle + O(\varepsilon^2)
\end{aligned}$$

and thus we obtain

$$\nu(\det B(\varepsilon)) = \nu(\det B(0)) + \varepsilon \nu'(\det B(0)) \det(B(0)) \langle \Delta, B(0)^{-1} \rangle + O(\varepsilon^2).$$

For simplicity let δ be

$$\delta = \det(B(0)) \langle \Delta, B(0)^{-1} \rangle \qquad (34)$$

then the equality

$$\nu(\det B(\varepsilon)) = \nu(\det B(0)) + \varepsilon \cdot \delta \cdot \nu'(\det B(0)) + O(\varepsilon^2) \qquad (35)$$

holds. By some calculation, we see that the asymptotic expansion of $B^{BFGS}[B_k; s + \varepsilon\bar{s},\ y + \varepsilon\bar{y}]$ and $(y+\varepsilon\bar{y})(y+\varepsilon\bar{y})^\top / (s+\varepsilon s)^\top (y+\varepsilon\bar{y})$ are respectively given by

$$\begin{aligned}
&B^{BFGS}[B_k; s + \varepsilon\bar{s}, y + \varepsilon\bar{y}]\\
&= B^{BFGS}[B_k; s, y]\\
&\quad + \varepsilon \left(\frac{y\bar{y}^\top + \bar{y}y^\top}{s^\top y} - \frac{s^\top \bar{y} + \bar{s}^\top y}{(s^\top y)^2} yy^\top - \frac{B_k (s\bar{s}^\top + \bar{s}s^\top) B_k}{s^\top B_k s} + \frac{2\bar{s}^\top B_k s}{(s^\top B_k s)^2} B_k s s^\top B_k \right)\\
&\quad + O(\varepsilon^2) \qquad (36)
\end{aligned}$$

and

$$\frac{(y+\varepsilon\bar{y})(y+\varepsilon\bar{y})^\top}{(s+\varepsilon\bar{s})^\top(y+\varepsilon\bar{y})} = \frac{yy^\top}{s^\top y} + \varepsilon\left(\frac{y\bar{y}^\top+\bar{y}y^\top}{s^\top y} - \frac{s^\top\bar{y}+\bar{s}^\top y}{(s^\top y)^2}yy^\top\right) + O(\varepsilon^2). \tag{37}$$

Substituting (35), (36) and (37) into the equality

$$B(\varepsilon) = \frac{\nu(\det B(\varepsilon))}{\nu(\det B_k)}B^{BFGS}[B_k; s+\varepsilon\bar{s}, y+\varepsilon\bar{y}] + \left(1-\frac{\nu(\det B(\varepsilon))}{\nu(\det B_k)}\right)\frac{(y+\varepsilon\bar{y})(y+\varepsilon\bar{y})^\top}{(s+\varepsilon\bar{s})^\top(y+\varepsilon\bar{y})},$$

we obtain

$$\begin{aligned}
&B(\varepsilon)\\
&= B(0) + \varepsilon\cdot\left\{\delta\cdot\frac{\nu'(\det B(0))}{\nu(\det B_k)}\big(B^{BFGS}[B_k;s,y] - \frac{yy^\top}{s^\top y}\big) + \frac{y\bar{y}^\top+\bar{y}y^\top}{s^\top y} - \frac{s^\top\bar{y}+\bar{s}^\top y}{(s^\top y)^2}yy^\top\right.\\
&\left. - \frac{\nu(\det B(0))}{\nu(\det B_k)}\frac{B_k(s\bar{s}^\top+\bar{s}s^\top)B_k}{s^\top B_k s} + \frac{\nu(\det B(0))}{\nu(\det B_k)}\frac{2\bar{s}^\top B_k s}{(s^\top B_k s)^2}B_k ss^\top B_k\right\} + O(\varepsilon^2),
\end{aligned}$$

and thus Δ is represented as

$$\begin{aligned}
\Delta &= \delta\cdot\frac{\nu'(\det B(0))}{\nu(\det B_k)}\left[B^{BFGS}[B_k;s,y] - \frac{yy^\top}{s^\top y}\right] + \frac{y\bar{y}^\top+\bar{y}y^\top}{s^\top y} - \frac{s^\top\bar{y}+\bar{s}^\top y}{(s^\top y)^2}yy^\top\\
&\quad - \frac{\nu(\det B(0))}{\nu(\det B_k)}\frac{B_k(s\bar{s}^\top+\bar{s}s^\top)B_k}{s^\top B_k s} + \frac{\nu(\det B(0))}{\nu(\det B_k)}\frac{2\bar{s}^\top B_k s}{(s^\top B_k s)^2}B_k ss^\top B_k\\
&= \delta\cdot\frac{\nu'(\det B(0))}{\nu(\det B(0))}\big[B(0) - \frac{yy^\top}{s^\top y}\big] + \frac{y\bar{y}^\top+\bar{y}y^\top}{s^\top y} - \frac{s^\top\bar{y}+\bar{s}^\top y}{(s^\top y)^2}yy^\top\\
&\quad - \frac{\nu(\det B(0))}{\nu(\det B_k)}\frac{B_k(s\bar{s}^\top+\bar{s}s^\top)B_k}{s^\top B_k s} + \frac{\nu(\det B(0))}{\nu(\det B_k)}\frac{2\bar{s}^\top B_k s}{(s^\top B_k s)^2}B_k ss^\top B_k
\end{aligned}$$

in which we use the equality

$$\frac{\nu(\det B(0))}{\nu(\det B_k)}\left[B^{BFGS}[B_k;s,y] - \frac{yy^\top}{s^\top y}\right] = B(0) - \frac{yy^\top}{s^\top y}.$$

Substituting the above Δ into (34), we have

$$\begin{aligned}
\delta = \frac{\det B(0)}{1-\beta(\det B(0))(n-1)}&\left\{\frac{s^\top\bar{y}-\bar{s}^\top y}{s^\top y}\right.\\
&\left. + \frac{\nu(\det B(0))}{\nu(\det B_k)}\left(\frac{2\bar{s}^\top B_k s\cdot s^\top B_k(B(0))^{-1}B_k s}{(s^\top B_k s)^2} - \frac{2\bar{s}^\top B_k(B(0))^{-1}B_k s}{s^\top B_k s}\right)\right\}.
\end{aligned}$$

As the result, we obtain

$$\begin{aligned}
&\frac{B(\varepsilon)-B(0)}{\varepsilon}\\
&=\left\{\frac{s^\top\bar{y}-\bar{s}^\top y}{s^\top y}+\frac{\nu(\det B(0))}{\nu(\det B_k)}\left(\frac{2\bar{s}^\top B_k s\cdot s^\top B_k(B(0))^{-1}B_k s}{(s^\top B_k s)^2}-\frac{2\bar{s}^\top B_k(B(0))^{-1}B_k s}{s^\top B_k s}\right)\right\}\\
&\times\frac{\beta(\det B(0))}{1-(n-1)\beta(\det B(0))}\left[B(0)-\frac{yy^\top}{s^\top y}\right]+\frac{y\bar{y}^\top+\bar{y}y^\top}{s^\top y}-\frac{s^\top\bar{y}+\bar{s}^\top y}{(s^\top y)^2}yy^\top\\
&+\frac{\nu(\det B(0))}{\nu(\det B_k)}\left[\frac{2\bar{s}^\top B_k s}{(s^\top B_k s)^2}B_k ss^\top B_k-\frac{B_k(s\bar{s}^\top+\bar{s}s^\top)B_k}{s^\top B_k s}\right]+O(\varepsilon).
\end{aligned}$$

Letting ε tend to zero, we obtain the influence function $\dot{B}(0)=\Delta[B_k;s,\bar{s},y,\bar{y}]$. □

Lemma 13. *Let $s,\bar{s},y$ and $\bar{y}$ be a set of column vectors in $\mathbb{R}^n$ such that $s^\top y>0$ and B_k be a matrix in $\mathrm{PD}(n)$. For an infinitesimal ε let $B(\varepsilon)$ be the optimal solution of*

$$\min_{B\in\mathrm{PD}(n)} D_V(B^{-1},B_k^{-1})\quad\text{subject to } B(s+\varepsilon\bar{s})=y+\varepsilon\bar{y}$$

and let $\Gamma[B_k;s,\bar{s},y,\bar{y}]$ be $\dot{B}(0)$ then we have

$$\Gamma[B_k;s,\bar{s},y,\bar{y}]=-B(0)\Delta[B_k^{-1};y,\bar{y},s,\bar{s}]B(0),\tag{38}$$

where Δ is the function defined in Lemma 12.

Proof. Let $H(\varepsilon)$ be the optimal solution of

$$\min_{H\in\mathrm{PD}(n)} D_V(H,B_k^{-1})\quad\text{subject to } H(y+\varepsilon\bar{y})=s+\varepsilon\bar{s}$$

then, clearly $B(\varepsilon)=H(\varepsilon)^{-1}$ holds. Thus we have

$$\Gamma[B_k;s,\bar{s},y,\bar{y}]=\dot{B}(0)=-H(0)^{-1}\dot{H}(0)H(0)^{-1}=-B(0)\Delta[B_k^{-1};y,\bar{y},s,\bar{s}]B(0),$$

where $\dot{H}(0)=\Delta[B_k^{-1};y,\bar{y},s,\bar{s}]$ is applied. □

We show another lemma which is useful to prove that the gross error sensitivity diverges to infinity.

Lemma 14. *Suppose $n\geq k+3$ for non-negative integers n and k. For any set of vectors $s,y,y_1\ldots,y_k\in\mathbb{R}^n$ such that $s^\top y>0$ and any positive real number d, there exists a sequence $\{B_i\}_{i=1}^\infty\subset\mathrm{PD}(n)$ satisfying the following three conditions:*

1. *The equalities* $B_i y = s$ *and* $B_i y_m = B_j y_m$ *hold for all* $i, j \geq 1$ *and* $m = 1, \ldots, k$.

2. $\det(B_i) = d$ *for all* $i \geq 1$.

3. $\lim_{i\to\infty} \|B_i\|_F = \infty$.

Proof. For any $s, y \in \mathbb{R}^n$ such that $s^\top y > 0$ there exists $\bar{B} \in \mathrm{PD}(n)$ satisfying $\bar{B}s = y$. Indeed, for the n by n identity matrix I, the matrix $\bar{B} = B^{BFGS}[I; s, y] \in \mathrm{PD}(n)$ is well-defined and satisfies $\bar{B}s = y$. When $n \geq k+3$ holds, there exist two unit vectors $p_1, p_2 \in \mathbb{R}^n$ satisfying $p_1^\top p_2 = 0$ and

$$\begin{aligned} &p_1^\top(\bar{B}^{1/2}s) = 0, \quad p_1^\top(\bar{B}^{1/2}y_m) = 0, \quad m = 1, \ldots, k, \\ &p_2^\top(\bar{B}^{1/2}s) = 0, \quad p_2^\top(\bar{B}^{1/2}y_m) = 0, \quad m = 1, \ldots, k. \end{aligned}$$

We will show that the matrix

$$B(a) = \bar{B}^{1/2}(I + ap_1p_1^\top + bp_2p_2^\top)\bar{B}^{1/2}$$

with

$$a > 0, \qquad b = \frac{d/\det(\bar{B})}{1+a} - 1 \tag{39}$$

satisfies four conditions: $B(a)s = y$, $B(a)y_m = \bar{B}y_m$, $\det B(a) = d$ and $B(a) \in \mathrm{PD}(n)$ for all $a > 0$. The first two equalities are clear from the definition of p_1, p_2 and $\bar{B}$. The determinant of $B(a)$ is equal to

$$\det(B(a)) = \det(\bar{B})\det(I + ap_1p_1^\top + bp_2p_2^\top) = \det(\bar{B})(1+a)(1+b) = d.$$

For any unit vector $x \in \mathbb{R}^n$ we have

$$\begin{aligned} x^\top(I + ap_1p_1^\top + bp_2p_2^\top)x &= 1 + a(p_1^\top x)^2 + b(p_2^\top x)^2 \\ &\geq 1 + b(p_2^\top x)^2 \qquad (\because a > 0) \\ &\geq 1 - (p_2^\top x)^2 \qquad (\because b > -1) \\ &\geq 0 \qquad \text{(Schwarz inequality)} \end{aligned}$$

and in addition the determinant of $(I + ap_1p_1^\top + bp_2p_2^\top)$ is equal to $d/\det(\bar{B}) > 0$. Thus $B(a)$ is positive definite. Let $\lambda_1(a)$ be the maximum eigenvalue of $B(a)$, and x be a unit vector defined by $x = \bar{B}^{-1/2}p_1/\|\bar{B}^{-1/2}p_1\|$. Then in terms of the maximum eigenvalue of $B(a)$ we have

$$\|B(a)\|_F \geq \lambda_1(a) \geq x^\top \bar{B}x + \frac{a}{p_1^\top \bar{B}^{-1}p_1}.$$

Then $\|B(a)\|_F$ tends to infinity when a tends to infinity. Thus the sequence defined by

$$B_i = B(i), \quad i = 1, 2, 3, \ldots \tag{40}$$

satisfies the conditions of the lemma. □

C.1 Proof of Theorem 7

Let $B(\varepsilon)$ be the optimal solution of (24). Under the inexact line search, the influence function $\dot{B}(0)$ for V-BFGS-B is equal to $\Delta[B_k; s, s, y, \bar{y}]$ which is defined in Lemma 12. Thus we have

$$\dot{B}(0) = \frac{(\bar{y}-y)^\top s}{s^\top y} \frac{\beta(\det B(0))}{1-(n-1)\beta(\det B(0))} \left[B(0) - \frac{yy^\top}{s^\top y} \right] + \frac{y\bar{y}^\top + \bar{y}y^\top}{s^\top y} - \frac{(y+\bar{y})^\top s}{(s^\top y)^2} yy^\top. \tag{41}$$

If $(\bar{y}-y)^\top s = 0$ holds for any $\bar{y} \in \mathcal{Y}$, the potential does not affect the norm of the influence function, because the first term of the above expression vanishes. Thus, clearly $V(z) = -\log(z)$ is an optimal potential. Below we assume $(\bar{y}-y)^\top s \neq 0$ for a vector $\bar{y} \in \mathcal{Y}$. Suppose that B_k satisfies $B_k s = y$. Then $B(0) = B_k$ holds, and the triangle inequality yields that

$$\begin{aligned}
\|\dot{B}(0)\|_F &= \|\Delta[B_k; s, s, y, \bar{y}]\|_F \\
&\geq \left| \frac{(\bar{y}-y)^\top s}{s^\top y} \right| \left| \frac{\beta(\det B_k)}{1-(n-1)\beta(\det B_k)} \right| \left(\|B_k\|_F - \left\| \frac{yy^\top}{s^\top y} \right\|_F \right) \\
&\quad - \left\| \frac{y\bar{y}^\top + \bar{y}y^\top}{s^\top y} - \frac{(\bar{y}+y)^\top s}{(s^\top y)^2} yy^\top \right\|_F.
\end{aligned}$$

If $\beta(z)$ is not the null function, there exists $d > 0$ such that $\beta(d) \neq 0$. Lemma 14 with $k = 0$ implies that for $n \geq 3$ there exists a sequence $\{\bar{B}_i\} \subset \mathrm{PD}(n)$ satisfying $\bar{B}_i s = y$, $\det \bar{B}_i = d$ for all i and $\lim_{i\to\infty} \|\bar{B}_i\|_F = \infty$. Hence

$$\lim_{i\to\infty} \|\Delta[\bar{B}_i; s, s, y, \bar{y}]\|_F = \infty$$

holds, and then we obtain

$$\sup\{\ \|\Delta[B_k; s, s, y, \bar{y}]\|_F \mid B_k \in \mathrm{PD}(n),\ \bar{y} \in \mathcal{Y}\ \} = \infty.$$

On the other hand, if $\beta(z) = 0$ for all $z > 0$, we obtain

$$\max_{B_k, \bar{y}} \|\Delta[B_k; s, s, y, \bar{y}]\|_F = \max_{\bar{y}\in\mathcal{Y}} \left\| \frac{y\bar{y}^\top + \bar{y}y^\top}{s^\top y} - \frac{(\bar{y}+y)^\top s}{(s^\top y)^2} yy^\top \right\|_F < \infty,$$

since $\mathcal{Y}$ is bounded. As the result, the potential V such that $\beta_V = 0$ minimizes the gross error sensitivity. The condition $\beta_V = 0$ leads to $V(z) = -\log(z)$ up to a constant factor.

C.2 Proof of Theorem 8

Let $B(\varepsilon)$ be the optimal solution of (25). Under the inexact line search, the influence function $\dot{B}(0)$ for V-DFP-B is equal to $\Gamma[B_k; s, s, y, \bar{y}]$ which is defined in Lemma 13.

First, we study the case that $\beta(z)$ is not the null function. For the matrix B_k such that $B_k s = y$, we have $B(0) = B_k$. Using Lemma 13 for $B(0) = B_k$, we have

$$\begin{aligned}\dot{B}(0) &= -B_k \Delta[B_k^{-1}; y, \bar{y}, s, s] B_k \\ &= \frac{(\bar{y}-y)^\top s}{s^\top y} \cdot \frac{\beta(\det(B_k)^{-1})}{1-(n-1)\beta(\det(B_k)^{-1})} \left[B_k - \frac{yy^\top}{s^\top y} \right] + \frac{y\bar{y}^\top + \bar{y}y^\top}{s^\top y} - \frac{(y+\bar{y})^\top s}{(s^\top y)^2} yy^\top,\end{aligned}$$

in which the equality $B_k s = y$ is used. The above expression is almost same as (41) with $B(0) = B_k$, and thus the same proof works to obtain

$$\sup\{ \ \|\dot{B}(0)\|_F \mid B_k \in \mathrm{PD}(n),\, \bar{y} \in \mathcal{Y} \ \} = \infty.$$

Next, we study the case that β is the null function, that is, $\beta(z) = 0$. Then, $V(z) = -\log(z)$ and $\nu(z) = 1$ hold. Let B_k be a positive definite matrix which does not necessarily satisfy $B_k s = y$. Then we obtain

$$\begin{aligned}\dot{B}(0) &= -B(0)\Delta[B_k^{-1}; y, \bar{y}, s, s] B(0) \\ &= -\frac{(y-\bar{y})^\top s}{(s^\top y)^2} yy^\top + \frac{B(0)B_k^{-1}(y\bar{y}^\top + \bar{y}y^\top)B_k^{-1}B(0)}{y^\top B_k^{-1} y} \\ &\quad - \frac{2\bar{y}^\top B_k^{-1} y}{(y^\top B_k^{-1} y)^2} B(0) B_k^{-1} yy^\top B_k^{-1} B(0)\end{aligned}$$

in which we used $B(0)s = y$. For $\beta = 0$, the updated matrix $B(0)$ is equal to $B^{DFP}[B_k; s, y]$ and thus, we have

$$B(0)B_k^{-1} = I - \frac{B_k s y^\top B_k^{-1} + y s^\top}{s^\top y} + \frac{s^\top B_k s}{(s^\top y)^2} yy^\top B_k^{-1} + \frac{1}{s^\top y} yy^\top B_k^{-1}. \quad (42)$$

Let $\bar{B} \in \mathrm{PD}(n)$ and c be a positive real number, and we define $t = \bar{B}s$, then for $B_k = c\bar{B}$ some calculation yields

$$\dot{B}(0) = -B(0)\Delta[(c\bar{B})^{-1}; y, \bar{y}, s, s]B(0) = -\frac{c}{s^\top y} Z - \frac{(y+\bar{y})^\top s}{(s^\top y)^2} yy^\top + \frac{y\bar{y}^\top + \bar{y}y^\top}{s^\top y},$$

where Z is defined by

$$Z = \left(t - \frac{s^\top t}{s^\top y} y \right) \left(\bar{y} - \frac{s^\top \bar{y}}{s^\top y} y \right)^\top + \left(\bar{y} - \frac{s^\top \bar{y}}{s^\top y} y \right) \left(t - \frac{s^\top t}{s^\top y} y \right)^\top.$$

Since $\mathcal{Y}$ contains an open subset, there exists a vector $\bar{y} \in \mathcal{Y}$ which is linearly independent to y. Clearly there exists $\bar{B} \in \mathrm{PD}(n)$ such that three vectors, $t = \bar{B}s$, $\bar{y}$ and y, are linearly independent. For such choice, Z is not the null matrix, and the equality

$$\lim_{c \to \infty} \|B(0)\Delta[(c\bar{B})^{-1}; y, \bar{y}, s, s]B(0)\|_F = \infty$$

holds. As the result, even for the standard DFP formula, we have

$$\sup\left\{\|\dot{B}(0)\|_F \mid B \in \mathrm{PD}(n),\, \bar{y} \in \mathcal{Y}\right\} = \infty.$$

In summary, for all V-DFP update for the Hessian approximation, the gross error sensitivity defined in Theorem 8 is equal to infinity.

C.3 Proof of Theorem 9

Let $H(\varepsilon)$ be the optimal solution of (26). Under the inexact line search, the influence function $\dot{H}(0)$ for V-BFGS-H is equal to $\Gamma[H_k; y, \bar{y}, s, s]$ which is defined in Lemma 13.

First, we study the case that $\beta(z)$ is not the null function. Suppose $\beta(d) \neq 0$. If H_k satisfies $H_k y = s$, then we have $H_k = H(0)$. Using Lemma 12 and Lemma 13 for the matrix H_k such that $H_k y = s$, we obtain

$$\begin{aligned}
&\dot{H}(0) \\
&= -H_k \Delta[H_k^{-1}; s, s, y, \bar{y}]H_k \\
&= \frac{(y-\bar{y})^\top s}{s^\top y} \frac{\beta(\det(H_k)^{-1})}{1-(n-1)\beta(\det(H_k)^{-1})}\left[H_k - \frac{ss^\top}{s^\top y}\right] - \frac{H_k \bar{y} s^\top + s\bar{y}^\top H_k}{s^\top y} + \frac{(y+\bar{y})^\top s}{(s^\top y)^2} ss^\top.
\end{aligned} \tag{43}$$

Lemma 14 with $k = 1$ implies that for $n \geq 4$ there exists a sequence $\{\bar{H}_i\} \subset \mathrm{PD}(n)$ satisfying the following conditions: $\bar{H}_i y = s$ and $(\det \bar{H}_i)^{-1} = d$ for all $i \geq 1$; $\bar{H}_i \bar{y} = \bar{H}_j \bar{y}$ for all $i, j \geq 1$; $\lim_{i \to \infty} \|\bar{H}_i\|_F = \infty$. We define $\bar{t} = \bar{H}_i \bar{y}$ which does not depend on i. Then for $H_k = \bar{H}_i$ we have

$$\begin{aligned}
&\|\dot{H}(0)\|_F \\
&= \|\bar{H}_i \Delta[\bar{H}_i^{-1}, s, s, y, \bar{y}]\bar{H}_i\|_F \\
&\geq \left|\frac{(y-\bar{y})^\top s}{s^\top y}\right| \left|\frac{\beta(d)}{1-(n-1)\beta(d)}\right| \left(\|\bar{H}_i\| - \left\|\frac{ss^\top}{s^\top y}\right\|\right) - \left\|\frac{\bar{t}s^\top + s\bar{t}^\top}{s^\top y} - \frac{(y+\bar{y})^\top s}{(s^\top y)^2} ss^\top\right\|.
\end{aligned}$$

Hence the equality

$$\lim_{i \to \infty} \|\bar{H}_i \Delta[\bar{H}_i^{-1}, s, s, y, \bar{y}]\bar{H}_i\|_F = \infty$$

holds, and thus we obtain

$$\sup\left\{ \|\dot{H}(0)\|_F \mid H_k \in \mathrm{PD}(n),\, \bar{y} \in \mathcal{Y} \right\} = \infty.$$

Next, we study the case that β is the null function, that is, $\beta(z) = 0$. Then, $V(z) = -\log(z)$ and $\nu(z) = 1$ holds. For H_k such that $H_k y = s$, we have

$$\dot{H}(0) = -\frac{H_k \bar{y} s^\top + s \bar{y}^\top H_k}{s^\top y} + \frac{(y+\bar{y})^\top s}{(s^\top y)^2} s s^\top. \tag{44}$$

Let $\bar{H}_0 \in \mathrm{PD}(n)$ be a matrix satisfying $\bar{H}_0 y = s$. Let $p_1 \in \mathbb{R}^n$ and $\bar{y} \in \mathcal{Y}$ be vectors satisfying $p_1^\top \bar{H}_0^{1/2} y = 0$ and $p_1^\top \bar{H}_0^{1/2} \bar{y} \neq 0$. For $n \geq 4$, the existence of p_1 and $\bar{y}$ is guaranteed by the assumption on $\mathcal{Y}$. Indeed, there exists $\bar{y} \in \mathcal{Y}$ such that $\bar{y}$ and y are linearly independent. We now define the matrix $\bar{H}_i \in \mathrm{PD}(n)$ by

$$\bar{H}_i = \bar{H}_0^{1/2}(I + i \cdot p_1 p_1^\top)\bar{H}_0^{1/2}, \quad i = 0, 1, 2, \ldots$$

Then we have

$$\bar{H}_i y = s, \quad \bar{H}_i \bar{y} = z + i \cdot u,$$

where $z = \bar{H}_0 \bar{y}$ and $u = (p_1^\top \bar{H}_0^{1/2} \bar{y}) \bar{H}_0^{1/2} p_1 \neq 0$. Substituting $H_k = \bar{H}_i$ into (44), we obtain

$$\dot{H}(0) = -i \cdot \frac{u s^\top + s u^\top}{s^\top y} + \frac{(y+\bar{y})^\top s}{s^\top y} s s^\top - \frac{z s^\top + s z^\top}{s^\top y}.$$

This implies that

$$\lim_{i \to \infty} \|\bar{H}_i \Delta[\bar{H}_i^{-1}; s, s, y, \bar{y}] \bar{H}_i\| = \infty.$$

for $\beta = 0$. Hence we obtain

$$\sup\left\{ \|\dot{H}(0)\|_F \mid H_k \in \mathrm{PD}(n),\, \bar{y} \in \mathcal{Y} \right\} = \infty$$

even for the standard BFGS update of the inverse Hessian approximation.

C.4 Proof of Theorem 10

Let $H(\varepsilon)$ be the optimal solution of (27). Under the inexact line search, the influence function $\dot{H}(0)$ for V-DFP-H is equal to $\Delta[H_k; y, \bar{y}, s, s]$ which is defined in Lemma 12.

First, we study the case that $\beta(z)$ is not the null function. Suppose $\beta(d) \neq 0$ for $d > 0$. If H_k satisfies $H_k y = s$, we have $H_k = H(0)$. Using Lemma 12 for the matrix H_k such that $H_k y = s$, we obtain

$$\begin{aligned}
&\dot{H}(0) \\
&= \Delta[H_k; y, \bar{y}, s, s] \\
&= \frac{(y-\bar{y})^\top s}{s^\top y} \frac{\beta(\det H_k)}{1-(n-1)\beta(\det H_k)} \left[H_k - \frac{ss^\top}{s^\top y} \right] - \frac{H_k \bar{y} s^\top + s\bar{y}^\top H_k}{s^\top y} + \frac{(y+\bar{y})^\top s}{(s^\top y)^2} ss^\top.
\end{aligned}$$

The above expression is almost same as (43), and thus the same proof remains valid to obtain

$$\sup \left\{ \|\dot{H}(0)\|_F \mid H_k \in \mathrm{PD}(n),\, \bar{y} \in \mathcal{Y} \right\} = \infty.$$

Next, we consider the case that β is the null function. Then $V(z) = -\log(z)$ and $\nu(z) = 1$ hold. For H_k such that $H_k y = s$, we have

$$\dot{H}(0) = \Delta[H_k; y, \bar{y}, s, s] = -\frac{H_k \bar{y} s^\top + s\bar{y}^\top H_k}{s^\top y} + \frac{(y+\bar{y})^\top s}{(s^\top y)^2} ss^\top.$$

This is the same as the influence function of (44), and thus, we obtain

$$\sup \left\{ \|\dot{H}(0)\|_F \mid H_k \in \mathrm{PD}(n),\, \bar{y} \in \mathcal{Y} \right\} = \infty.$$

References

[1] S. Amari and H. Nagaoka. *Methods of Information Geometry*, volume 191 of *Translations of Mathematical Monographs*. Oxford University Press, 2000.

[2] A. Banerjee, S. Merugu, I. Dhillon, and J. Ghosh. Clustering with bregman divergences. *Journal of Machine Learning Research*, 6:1705–1749, 2005.

[3] L. M. Bregman. The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming. *USSR Computational Mathematics and Mathematical Physics*, 7:200–217, 1967.

[4] C. G. Broyden. Quasi-newton methods and their application to function minimisation. *Mathematics of Computation*, 21(99):368–381, 1967.

[5] A. R. Conn, N. I. M Gould, and P. L. Toint. Testing a class of algorithms for solving minimization problems with simple bounds on the variables. *Mathematics of Computation*, 50:399–430, 1988.

[6] I. Dhillon and J. Tropp. Matrix nearness problems with bregman divergences. *SIAM J. Matrix Anal. Appl.*, 29(4):1120–1146, 2007.

[7] R. Fletcher. A new result for quasi-Newton formulae. *SIMA J. Optim.*, 1:18–21, 1991.

[8] R. Fletcher. An optimal positive definite update form sparse hessian matrices. *SIMA J. Optim.*, 5:192–218, 1995.

[9] P. E. Gill and W. Murray. Quasi-Newton methods for unconstrained optimization. *J. Inst. Math Appl.*, 9:91–108, 1972.

[10] O. Güler, F. Gürtuna, and O. Shevchenko. Duality in quasi-newton methods and new variational characterizations of the DFP and BFGS updates. *Optimization Methods and Software*, 24(1):45–62, 2009.

[11] F. R. Hampel, P. J. Rousseeuw, E. M. Ronchetti, and W. A. Stahel. *Robust Statistics. The Approach based on Influence Functions.* John Wiley and Sons, Inc., 1986.

[12] S. Kullback and R. A. Leibler. On information and sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86, 1951.

[13] D. Luenberger and Y. Ye. *Linear and Nonlinear Programming.* Springer, 2008.

[14] N. Murata, T. Takenouchi, T. Kanamori, and S. Eguchi. Information geometry of U-Boost and Bregman divergence. *Neural Computation*, 16(7):1437–1481, 2004.

[15] J. Nocedal and S. Wright. *Numerical Optimization.* Springer, 1999.

[16] J. Nocedal and Y.-X. Yuan. Analysis of a self-scaling quasi-newton method. *Math. Program.*, 61:19–37, 1993.

[17] A. Ohara and S. Eguchi. Geometry on positive definite matrices and v-potential function. Technical report, ISM Research Memo, 2005.

[18] S. S. Oren and D. G. Luenberger. Self-scaling variable metric (ssvm) algorithms, part i. criteria and sufficient conditions for scaling a class of algorithms. *Management Science*, 20:845–862, 1974.

[19] M. J. D. Powell. How bad are the bfgs and dfp methods when the objective function is quadratic? *Math. Prog.*, 34(1):34–47, 1986.

[20] N. Yamashita. Sparse quasi-newton updates with positive definite matrix completion. *Math. Program.*, 115(1):1–30, 2008.

Table 2: Approximate influence function for V-BFGS update and V-DFP update is shown. The power potential $V(z) = (1 - z^\gamma)/\gamma$ is used for V-extended quasi-Newton methods, where $\gamma = 0$ corresponds to BFGS or DFP method.

V-BFGS-B

B_k	$\mathrm{diag}(1,\dots,n)/(n!)^{1/n}$			$\mathrm{diag}(1,\dots,n)$			$I + n^3 pp^\top$		
γ	-2	-1	0	-2	-1	0	-2	-1	0
$n = 10$	9.5e+00	9.5e+00	9.5e+00	1.5e+01	9.7e+00	9.5e+00	2.0e+02	1.0e+02	5.0e+01
$n = 100$	2.7e+01	2.7e+01	2.7e+01	2.3e+02	2.8e+01	2.7e+01	1.1e+04	1.0e+04	8.7e+03
$n = 500$	9.3e+01	9.3e+01	9.3e+01	2.8e+03	9.6e+01	9.3e+01	2.6e+05	2.5e+05	2.4e+05
$n = 1000$	1.0e+02	1.0e+02	1.0e+02	7.4e+03	1.1e+02	1.0e+02	1.0e+06	9.9e+05	9.7e+05

V-DFP-B

B_k	$\mathrm{diag}(1,\dots,n)/(n!)^{1/n}$			$\mathrm{diag}(1,\dots,n)$			$I + n^3 pp^\top$		
γ	-2	-1	0	-2	-1	0	-2	-1	0
$n = 10$	1.3e+02	1.3e+02	1.3e+02	2.9e+03	6.5e+02	1.5e+02	2.0e+02	1.0e+02	5.0e+01
$n = 100$	1.7e+03	1.7e+03	1.7e+03	2.5e+06	6.5e+04	1.7e+03	1.1e+04	1.0e+04	8.7e+03
$n = 500$	4.6e+04	4.6e+04	4.6e+04	1.6e+09	8.7e+06	4.7e+04	2.6e+05	2.5e+05	2.4e+05
$n = 1000$	3.0e+04	3.0e+04	3.0e+04	4.1e+09	1.1e+07	3.0e+04	1.0e+06	9.9e+05	9.7e+05

V-BFGS-H

H_k	$\mathrm{diag}(1,\dots,n)/(n!)^{1/n}$			$\mathrm{diag}(1,\dots,n)$			$I + n^3 pp^\top$		
γ	-2	-1	0	-2	-1	0	-2	-1	0
$n = 10$	2.1e+02	2.1e+02	2.1e+02	4.8e+03	1.1e+03	2.4e+02	2.2e+02	1.1e+02	5.6e+01
$n = 100$	1.1e+03	1.1e+03	1.1e+03	1.6e+06	4.1e+04	1.1e+03	2.0e+04	1.7e+04	1.5e+04
$n = 500$	8.2e+04	8.2e+04	8.2e+04	2.8e+09	1.5e+07	8.3e+04	8.7e+05	8.4e+05	8.1e+05
$n = 1000$	2.6e+04	2.6e+04	2.6e+04	3.6e+09	9.8e+06	2.7e+04	4.7e+06	4.6e+06	4.5e+06

V-DFP-H

H_k	$\mathrm{diag}(1,\dots,n)/(n!)^{1/n}$			$\mathrm{diag}(1,\dots,n)$			$I + n^3 pp^\top$		
γ	-2	-1	0	-2	-1	0	-2	-1	0
$n = 10$	1.0e+01	1.0e+01	1.0e+01	1.7e+01	1.1e+01	1.0e+01	2.5e+02	1.3e+02	6.4e+01
$n = 100$	2.1e+01	2.1e+01	2.1e+01	4.5e+02	2.5e+01	2.1e+01	4.1e+06	3.6e+06	3.1e+06
$n = 500$	9.9e+01	9.9e+01	9.9e+01	9.5e+03	1.2e+02	9.9e+01	1.4e+09	1.4e+09	1.3e+09
$n = 1000$	1.2e+02	1.2e+02	1.2e+02	3.6e+04	1.7e+02	1.2e+02	1.2e+10	1.2e+10	1.2e+10

Table 3: Number of iterations by BFGS and DFP under inexact line search. The number of h denotes intensity of noise involved in the line search.

		$n = 100$		$n = 500$		$n = 1000$	
	h	BFGS	DFP	BFGS	DFP	BFGS	DFP
Problem 1	0.0	100.4	110.6	434.6	577.8	682.1	1788.5
	0.1	102.9	166.2	430.6	1165.2	680.9	2628.9
	0.2	104.5	198.6	443.6	1361.8	685.1	3099.2
	0.3	106.0	223.0	444.2	1501.6	687.6	3365.9
Problem 2	0.0	100.9	111.6	428.5	585.7	661.5	2489.8
	0.1	102.8	153.5	443.5	1237.4	672.4	2762.1
	0.2	104.4	177.7	438.3	1419.6	682.7	3301.2
	0.3	106.1	199.4	454.0	1592.8	694.0	3730.8